EURANDOM PREPRINT SERIES

2016-011

October 14, 2016

Queue-length balance equations in multiclass multiserver queues and their generalizations

M. Boon, O. Boxma, O. Kella, M. Miyazawa
ISSN 1389-2355

Queue-length balance equations in multiclass multiserver queues and their generalizations

Marko A.A. Boon[*] Onno J. Boxma [†‡] Offer Kella[§¶] Masakiyo Miyazawa[‖**]

October 14, 2016

Abstract

A classical result for the steady-state queue-length distribution of single-class queueing systems is the following: the distribution of the queue length just before an arrival epoch equals the distribution of the queue length just after a departure epoch. The constraint for this result to be valid is that arrivals, and also service completions, with probability one occur individually, i.e., not in batches.

In the first half of this paper, we show that it is easy to write down somewhat similar balance equations for multidimensional queue-length processes for a large family of multiclass multiserver queues with Poisson arrivals – even when arrivals may occur in batches. We demonstrate the use of these balance equations, in combination with PASTA, by (i) providing very simple derivations of some known results for polling systems, and (ii) obtaining new results for some queueing systems with priorities.

In the second half of the paper, we formally verify those balance equations under a general framework. They are called distributional relationships, and are obtained for any external arrival process and state dependent routing as long as certain stationarity conditions are satisfied and external arrivals and service completions do not simultaneously occur. We also extend the distributional relationships for a non-stationary framework.

Keywords: queue length; steady-state distribution; balance equations; distributional relationship; Palm distribution; non-stationary framework.

1 Introduction

A classical result for the steady-state queue-length distribution of single-class queueing systems is the following: the distribution of the queue length just before an arrival epoch equals

[*]Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands. m.a.a.boon@tue.nl

[†]Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands. o.j.boxma@tue.nl

[‡]Partly funded by the NWO Gravity Project NETWORKS, Grant Number 024.002.003.

[§]Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel. offer.kella@gmail.com

[¶]Supported in part by grant 1462/13 from the Israel Science Foundation and the Vigevani Chair in Statistics.

[‖]Department of Information Sciences, Tokyo University of Science, Noda City, Chiba 278, Japan. miyazawa@rs.tus.ac.jp

[**]Supported in part by JSPS KAKENHI Grant Number 16H027860001.

the distribution of the queue length just after a departure epoch. The constraint for this result to be valid is that, with probability one, arrivals, and also service completions, occur individually, i.e., not in batches. The result then follows by a simple level-crossing argument: in steady state, the event that a customer arrives to find j customers present occurs just as often as the event that a customer leaves j customers behind, for all $j = 0, 1, \ldots$.

At first sight this level-crossing argument breaks down in higher dimensions, like in the case of multiple customer classes. Indeed, with $\boldsymbol{x} \geq \mathbf{0}$ and $\boldsymbol{e}_k$ being a unit vector with 1 in the kth coordinate and zero elsewhere, an m-dimensional process can leave state $\boldsymbol{x}$ because of an arrival of a customer of type i, and enter that state from state $\boldsymbol{x} + \boldsymbol{e}_k$ because of a departure of a customer of *another* type k. However, we shall argue that it is easy to write down a more global balance equation for multidimensional queue length processes for a large class of queues and queueing networks – also when service times are not exponentially distributed, and even when arrivals may occur in batches. Subsequently we shall explore that fact to obtain a simple relation between the steady-state joint queue-length distribution at arrival epochs (which under various circumstances is equal to the time average distribution) and at service completion epochs. Once one has a relation between the probability generating function (PGF) at arbitrary epochs and at service completion epochs, one can find the former when one has the latter. The latter results are indeed known in an $M/G/1$ setting, where it is natural to look at departure epochs. This will yield both new results (for multiclass queueing models with fixed priorities and for the longer-queue model), as well as new and simple derivations of known results for, e.g., polling models. In fact, the research for the present paper was initially motivated by the desire to provide an intuitive explanation of a result in [3] regarding the steady-state joint queue-length distribution in a large class of polling models.

Although balance equations are intuitively appealing, their mathematical verification may require a large amount of work. This motivates to derive those distributional relationships in a unified way using a general tool. The so called rate conservation law is such a tool as demonstrated in [10] (also see [1, 9]). This method is applicable to a general model, but requires Palm distributions, which may not be easy to understand.

In Section 6 and Section 7 of this paper we take yet another approach, based on a time evolution of a sample path. This approach is parallel to the rate conservation law, but does not require Palm distributions, which are replaced by sample averages. We apply it to a general model, and derive a distributional relationship among different embedded epochs. All the results which are obtained from balance equations in Sections 2–5 can be formally obtained as its special cases. Furthermore, a non-stationary version of the distributional relationship is derived with some error term, which vanishes as time goes to infinity.

The paper is organized as follows. Section 2 provides a short proof of Theorem 1 of [3] by using a multi-dimensional queue-length balance argument. Section 3 presents a more general form of the balance argument and the ensuing relation between the steady-state joint queue-length distribution at arbitrary epochs and at departure epochs. An extension to networks is given in Section 4. Some applications are discussed in Section 5. Section 6 derives the distributional relationship for an open queueing network under a very general setting in Theorem 3. Its corollaries cover major results in Sections 2–5. Extensions of Theorem 3 and Corollary 1 are discussed in Section 7. Concluding remarks are given in Section 8.

2 A balance equation for a class of polling models

In this section we provide a simple relation between the steady-state joint queue-length distribution at arbitrary epochs and at departure epochs for polling models. This relation, which is derived by introducing a multi-dimensional queue-length balance argument, is used to provide a short, but somewhat intuitive derivation of Theorem 1 of [3]. In the next sections we shall generalize and give more rigorous derivations for that balance equation. Let us first describe the polling model studied in [3].

Consider a system of $m \geq 1$ infinite-buffer queues $Q_1, \ldots, Q_m$ and a single server S. Queues are indexed by $J = \{1, 2, \ldots, m\}$. The service times of customers in Q_i are i.i.d. (independent, identically distributed) positive random variables generically denoted by B_i, with means $b_i := \mathbb{E}B_i$. Denote the Laplace-Stieltjes transform (LST) of B_i by $\tilde{B}_i(\cdot)$. The server moves among the queues in a cyclic order. When S moves from Q_i to Q_{i+1}, it incurs a switchover period. The durations of successive switchover times are i.i.d. non-negative random variables, which we generically denote by S_i. Denote the LST of S_i by $\tilde{S}_i(\cdot)$ and assume that $s_i := \mathbb{E}S_i < \infty$; let $s := \sum_{i=1}^m s_i$. Customers arrive at Q_i according to a Poisson process with rate λ_i; let $\lambda := \sum_{i=1}^m \lambda_i$. We do not assume anything about the service disciplines at Q_i. Define $\rho_i := \lambda_i b_i$ as the traffic intensity at Q_i; let $\rho := \sum_{i=1}^m \rho_i$. We assume that $\rho < 1$, which is a necessary condition for the system to be stable. In what follows we shall write $\boldsymbol{z}$ for an m-dimensional vector in $\mathbb{R}^m$, $\boldsymbol{z} = (z_1, \ldots, z_m)$, and we assume that $|z_i| \leq 1$ for every $i \in J$. We implicitly use the convention that any index summation is modulo m, for example $Q_{m+1} \equiv Q_1$.

Assume that all the usual independence assumptions hold between the service times, the switchover times and the interarrival times. We assume that the ergodicity conditions are fulfilled and we restrict ourselves to results for the stationary situation.

Now introduce the PGF of various joint queue-length distributions: $V_i^b(\boldsymbol{z})$ and $V_i^c(\boldsymbol{z})$ denote the PGF's of the joint queue-length distribution at visit beginnings and visit completions at Q_i, while $S_i^b(\boldsymbol{z})$ and $S_i^c(\boldsymbol{z})$ denote the PGF's of the joint queue-length distribution at service beginnings and service completions at Q_i; $L(\boldsymbol{z})$ denotes the PGF of the joint queue-length distribution at an arbitrary time in steady-state. Theorem 1 of [3] states that, with mean cycle time $\mathbb{E}C = \frac{s}{1-\rho}$:

$$L(\boldsymbol{z}) = \frac{1}{\mathbb{E}C} \sum_{i=1}^{m} \left(\frac{V_i^b(\boldsymbol{z}) - V_i^c(\boldsymbol{z})}{\Sigma(\boldsymbol{z})} \frac{z_i \left(1 - \tilde{B}_i(\Sigma(\boldsymbol{z}))\right)}{z_i - \tilde{B}_i(\Sigma(\boldsymbol{z}))} + \frac{V_i^c(\boldsymbol{z}) - V_{i+1}^b(\boldsymbol{z})}{\Sigma(\boldsymbol{z})} \right), \tag{1}$$

with $\Sigma(\boldsymbol{z}) := \sum_{j=1}^m \lambda_j (1 - z_j)$.

Its proof in [3] is based on the following relations:
(i) a balance relation for polling systems, which is due to Eisenberg [7] and which was generalized in [2]:

$$\gamma_i V_i^b(\boldsymbol{z}) + S_i^c(\boldsymbol{z}) = S_i^b(\boldsymbol{z}) + \gamma_i V_i^c(\boldsymbol{z}), \quad i \in J. \tag{2}$$

Here $\gamma_i := 1/\lambda_i \mathbb{E}C$ represents the reciprocal of the mean number of customers served at Q_i per visit, i.e., the long-term ratio of visit beginnings to service beginnings.
(ii) an obvious relation between queue lengths at the beginning and end of a service time:

$$S_i^c(\boldsymbol{z}) = S_i^b(\boldsymbol{z}) \frac{\tilde{B}_i(\Sigma(\boldsymbol{z}))}{z_i}, \quad i \in J. \tag{3}$$

(iii) an obvious relation between queue lengths at the beginning and end of a switchover time:

$$V_{i+1}^b(\boldsymbol{z}) = V_i^c(\boldsymbol{z})\tilde{S}_i\left(\Sigma(\boldsymbol{z})\right), \quad i \in J. \tag{4}$$

(iv) a stochastic mean value theorem, expressing $L(\boldsymbol{z})$ as an average over the PGFs of the joint queue-length distribution at an arbitrary moment during a visit to Q_i ($X_i(\boldsymbol{z})$) and during a switchover period between Q_i and Q_{i+1} ($Y_i(\boldsymbol{z})$):

$$L(\boldsymbol{z}) = \frac{1}{\mathbb{E}C}\sum_{i=1}^{m}\left(\frac{b_i}{\gamma_i}X_i(\boldsymbol{z}) + s_i Y_i(\boldsymbol{z})\right), \tag{5}$$

where, for $i \in J$,

$$X_i(\boldsymbol{z}) = S_i^b(\boldsymbol{z})\tilde{B}_i^{\text{past}}(\Sigma(\boldsymbol{z})), \tag{6}$$

where $\tilde{B}_i^{\text{past}}(\cdot)$ is the LST of B_i^{past}, the past part of B_i, and

$$Y_i(\boldsymbol{z}) = V_i^c(\boldsymbol{z})\tilde{S}_i^{\text{past}}(\Sigma(\boldsymbol{z})), \tag{7}$$

where $\tilde{S}_i^{\text{past}}(\cdot)$ is the LST of S_i^{past}, the past part of S_i. Starting from (5), substituting (6) and (7), and using (2) and (3) to eliminate all $S_i^c(\boldsymbol{z})$ and $S_i^b(\boldsymbol{z})$, yields (1).

Remark 1 In [3] also zero switchover times are allowed; the same result (1) is shown to hold.

In Theorem 1 of [3] it was subsequently observed that one may simplify (1) as follows, by using (2) and (3):

$$L(\boldsymbol{z}) = \frac{\sum_{i=1}^m \lambda_i(1-z_i)S_i^c(\boldsymbol{z})}{\sum_{i=1}^m \lambda_i(1-z_i)}. \tag{8}$$

This formula is remarkably simple; please notice that it does not involve the service time distributions, and that the service disciplines at the various queues also do not play a role, which suggests that (1) is based on very general principles. This is the formula for which we would like to provide a short proof. In combination with (2) – (4), it also gives a short proof of (1). In other words, one can obtain an expression for the PGF of the joint steady-state queue-length distribution in a large class of polling systems by just using the elementary balance equations (2) and (9) (see below), combined with the obvious relations (3) and (4).

Short proof of (8).
First rewrite (8) into

$$\sum_{i=1}^{m}\lambda_i(1-z_i)L(\boldsymbol{z}) = \sum_{i=1}^{m}\lambda_i(1-z_i)S_i^c(\boldsymbol{z}). \tag{9}$$

Secondly, observe that, because of the Poisson arrival processes, $L(\boldsymbol{z})$ also is the PGF of the joint queue-length distribution just before an arrival at Q_i, $i \in J$ by PASTA (Poisson Arrival See Time Averages, e.g., see [1, 9]).
Thirdly, invert the transform expressions on both sides of (9), yielding for $\boldsymbol{x} \geq \boldsymbol{0}$ and $\boldsymbol{e}_i$ being the unit vector with 1 in the ith coordinate and zero elsewhere:

$$\sum_{i=1}^{m}\lambda_i\pi_i^e(\boldsymbol{x}) - \sum_{i=1}^{m}\lambda_i\pi_i^e(\boldsymbol{x}-\boldsymbol{e}_i) = \sum_{i=1}^{m}\lambda_i\pi_i^d(\boldsymbol{x}) - \sum_{i=1}^{m}\lambda_i\pi_i^d(\boldsymbol{x}-\boldsymbol{e}_i), \tag{10}$$

where $\pi_i^d(\cdot)$ indicates that we consider the joint queue-length distribution right *after* a departure from Q_i, and $\pi_i^e(\cdot)$ denotes that we view the system just *before* an external arrival at Q_i. Fourthly, we reshuffle the terms:

$$\sum_{i=1}^{m} \lambda_i \pi_i^e(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \pi_i^d(\boldsymbol{x} - \boldsymbol{e}_i) = \sum_{i=1}^{m} \lambda_i \pi_i^d(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \pi_i^e(\boldsymbol{x} - \boldsymbol{e}_i). \tag{11}$$

Finally, observe that the lefthand side of (11) represents the rate out of state $\boldsymbol{x}$, and the righthand side represents the rate into that state. Indeed, the first term in the lefthand side corresponds to arrivals which find $\boldsymbol{x}$ customers in the system. The second term in the lefthand side is slightly less obvious. It corresponds to departures that take place in state $\boldsymbol{x}$. Notice that the rate at which customers depart from Q_i equals λ_i (although the departure process will not be a Poisson process), and that $\pi_i^d(\boldsymbol{x} - \boldsymbol{e}_i)$ is the fraction of departures from Q_i which take the system out of state $\boldsymbol{x}$. Similarly interpret the terms in the righthand side. We conclude that (8) amounts to a simple flow balance formula.

Remark 2 A similar flow balance argument was used in [5] to derive a queue-length expression in an $M/G/1$ FCFS queue with multiple customer classes.

Remark 3 Observe that (8) immediately gives the formula for the marginal distributions. Indeed, for a vector $\boldsymbol{z}_{m,i} = (1, \ldots, 1, z_i, 1, \ldots, 1)$, $L(\boldsymbol{z}_{m,i}) = S_i^c(\boldsymbol{z}_{m,i})$. From the well-known 'step' (level-crossing) argument it follows that $S_i^c(\boldsymbol{z}_{m,i})$ is also the PGF of the queue-length distribution in Q_i at an *arrival* epoch at Q_i. By PASTA it is also the PGF of the steady-state distribution of Q_i.

Next take $\boldsymbol{z}_T = (z, \ldots, z)$. (8) now states that the PGF of the distribution of the total queue length (in terms of z) equals $\sum_{i=1}^m \lambda_i S_i^c(\boldsymbol{z}_T) / \sum_{j=1}^m \lambda_j$. This formula may be interpreted as follows. By PASTA, $L(\boldsymbol{z}_T)$ is also the PGF of the distribution of the total queue length at an arrival epoch. By a level-crossing argument, it follows that this equals the PGF of the distribution of the total queue length just after a departure epoch. The result now follows from the observation that a fraction $\lambda_i / \sum_{j=1}^m \lambda_j$ of the departure epochs refers to a departure from Q_i.

Remark 4 Relation (8) may be viewed as an m-dimensional version of the above-mentioned one-dimensional 'step' (level-crossing) relation that holds for queues with single arrivals and single departures.

3 Formulation and proof of the general result for single-node systems

In this section we consider an m-class single-node service facility, with $m \geq 1$. We allow multiple servers. Customers arrive according to a Poisson process, possibly in batches. Customers of class i require service at the service facility according to service time distribution $B_i(\cdot)$, $i \in J$. These distributions are assumed to be continuous, but not otherwise specified. No customers are lost; there is an infinite waiting room (see a remark at the end of Section 6 for a finite buffer case). After completion of their service, customers immediately leave. We assume that the steady-state joint queue-length distribution (numbers of customers of all classes in the system) exists. As before, its PGF is denoted by $L(\boldsymbol{z})$. We also again denote the

PGF of the steady-state joint queue-length distributions immediately after departure epochs of a class i-customer by $S_i^c(\boldsymbol{z})$, $i \in J$. We do not specify according to which service discipline the customers are served; polling with FCFS within each class is just one of many options.

In this section we formulate and prove two theorems. Theorem 1 generalizes (8) to the above-described m-class single-node multi-server service facility with batch arrivals. Theorem 2 provides a further generalization to batch services, with a fixed batch size.

Theorem 1 *Consider the above-described m-class single-node service facility. Assume that customers arrive according to a batch Poisson process with rate λ and that customers are served individually, in some non-specified order. Let an arbitrary batch arrival have size $\boldsymbol{G} = (G_1, \ldots, G_m)$ with distribution $q(i_1, \ldots, i_m)$ and PGF $\mathbb{E}[\boldsymbol{z}^{\boldsymbol{G}}] = \mathbb{E}[z_1^{G_1} \ldots z_m^{G_m}]$. Then the following relation holds between the PGF $L(\boldsymbol{z})$ and the PGFs $S_i^c(\boldsymbol{z})$, $i \in J$:*

$$(1 - \mathbb{E}[\boldsymbol{z}^{\boldsymbol{G}}])L(\boldsymbol{z}) = \sum_{i=1}^{m} (1 - z_i)\mathbb{E}G_i S_i^c(\boldsymbol{z}). \tag{12}$$

Proof Starting-point is the observation that there is equality between the rate out of state $\boldsymbol{j} \equiv (j_1, \ldots, j_i, \ldots, j_m)$ and into state $\boldsymbol{j}$. This leads to the following balance equation, which is similar to (11) (denote by A that we observe the system just before a batch arrives): for $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$,

$$\begin{aligned} &\lambda \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|A) + \sum_{i=1}^{m} \lambda \mathbb{E}G_i \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_i|D_i) \\ &= \sum_{i=1}^{m} \lambda \mathbb{E}G_i \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|D_i) + \lambda \sum_{i_1 \le j_1} \cdots \sum_{i_m \le j_m} q(\boldsymbol{i}) \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{i}|A). \end{aligned} \tag{13}$$

The required rigor to this balance argument will appear later. Notice that $\lambda \mathbb{E}G_i$ is the rate at which departures of class-i customers occur. Further notice that all terms can be divided by λ. Now swap the last term of the lefthand side and of the righthand side, and take generating functions. This yields (12), after observing that PASTA holds (of course, we could have replaced $L(\boldsymbol{z})$ in (12) by the PGF of the joint queue-length distribution just before a batch arrival). □

Remark 5 Special cases of the above theorem are obtained by assuming that batches always contain only customers of one type. For the special case that batches have just one customer of class i with probability $\frac{\lambda_i}{\lambda}$, $i \in J$, (12) reduces to (8) that was obtained for a polling system (but that obviously holds for a much more general class of service disciplines).

Let us now generalize Theorem 1 by allowing a, rather restricted, form of batch service. See Remarks 6 and 7 for a discussion why the possibility of simultaneous departures, generally speaking, leads to serious complications.

Theorem 2 *Consider the m-class single-node service facility of Theorem 1, with the additional assumption that customers of class-i are always served in batches of fixed size K_i, $i \in J$; the start of a service of class-i customers is delayed until K_i customers are present. Then the following relation holds between the PGF $L(\boldsymbol{z})$ and the PGFs $S_i^c(\boldsymbol{z})$, $i \in J$:*

$$(1 - \mathbb{E}[\boldsymbol{z}^{\boldsymbol{G}}])L(\boldsymbol{z}) = \sum_{i=1}^{m} \frac{1 - z_i^{K_i}}{K_i} \mathbb{E}G_i S_i^c(\boldsymbol{z}). \tag{14}$$

Proof Compared to the proof of Theorem 1, the only changes occur in the terms with condition D_i. The rate at which there is a departure of class-i batches is $\frac{\lambda \mathbb{E}G_i}{K_i}$. Furthermore, when state $(j_1, \ldots, j_i, \ldots, j_m)$ is left because of a departure of a class-i batch, the next state always is $(j_1, \ldots, j_i - K_i, \ldots, j_m)$ (with $j_i \geq K_i$). □

Remark 6 In the case of batch service with non-fixed size, we face the following problem. Suppose that a batch service reaches completion when the system is in state $(j_1, \ldots, j_m)$. It is then not a priori clear to which state a transition takes place; this may depend on $(j_1, \ldots, j_m)$ in a complicated way, which seems to destroy the simplicity of our balance equations.

Remark 7 The reason why in this section we have required continuous service time distributions is similar to the reason for restricting ourselves in Theorem 2 to batches of a fixed size. Without this restriction, the following may happen. Suppose that a batch of jobs arrives in which two jobs have, with positive probability, the same service time requirement. Upon their arrival, there might be two idle servers who simultaneously start their service – and simultaneously end their service.

Remark 8 We also have restricted ourselves to individual service. Otherwise one could think of disciplines like: longest-customer-first processor sharing. There it is possible that a server serves several customers simultaneously, and that all their remaining service times are the same. Or one could have a service discipline which uses very specific knowledge; e.g., server 1 delays service until the remaining service time at server 2 exactly equals the service time of the customer he is going to serve. In such cases, several customers may simultaneously leave a queue in a way that is a priori difficult to predict.

The problems in these remarks will be shown to be a matter of formulation in Section 6. We there resolve them in a certain sense under a general framework.

4 A simple network

We first consider a simple network of two queues, to get insight into the following question: To which extent can our approach of equating flows at arrival and departure epochs, and using them to get relations between PGFs, be extended to the case of a network with internal movements of customers? In the latter case, a departure from one queue may coincide with an arrival at another (or the same) queue. The network under consideration consists of two queues Q_1 and Q_2, with two independent external Poisson arrival processes, with rate λ_i at Q_i, $i = 1, 2$. Q_1 and Q_2 are service facilities with continuous, otherwise non-specified, service time distributions. After service completion at Q_1, customers always move to Q_2. As before, D_i denotes that we view the system just after a departure from Q_i that leaves the system, and A_i denotes that we view the system just before an external arrival at Q_i. In addition, I_{ij} denotes that we view the system just after a departure from Q_i and just before the arrival of that customer at Q_j. We then can write down the following flow balancing equations for $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{j} = (j_1, j_2)$.

$$\begin{aligned}\lambda_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|A_1) + \lambda_2 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|A_2) + (\lambda_1 + \lambda_2)\mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_2|D_2) + \lambda_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_1|I_{12}) \\ = \lambda_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_1|A_1) + \lambda_2 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_2|A_2) \\ + (\lambda_1 + \lambda_2)\mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|D_2) + \lambda_1 \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_2|I_{12}). \qquad (15)\end{aligned}$$

Taking PGFs and using PASTA results in:

$$\sum_{i=1}^{2} \lambda_i (1 - z_i) L(\boldsymbol{z}) = (\lambda_1 + \lambda_2)(1 - z_2) S_2^c(\boldsymbol{z}) + \lambda_1 (z_2 - z_1) T_{12}(\boldsymbol{z}), \tag{16}$$

where $T_{12}(\cdot)$ denotes the PGF of the joint queue-length distribution just after a customer has departed from Q_1 and just before it has arrived at Q_2. As long as a customer's routing does not depend on the state of the system, $T_{12}(\boldsymbol{z}) \equiv S_1^c(\boldsymbol{z})$, resulting in

$$\sum_{i=1}^{2} \lambda_i (1 - z_i) L(\boldsymbol{z}) = \lambda_1 (z_2 - z_1) S_1^c(\boldsymbol{z}) + (\lambda_1 + \lambda_2)(1 - z_2) S_2^c(\boldsymbol{z}). \tag{17}$$

Putting $z_1 = z_2 = z$ gives $L(z, z) = S_2^c(z, z)$, confirming equality of the total queue-length distribution just before an arrival to the system (use PASTA) and just after a departure from the system. Notice that in a Jackson network $S_1^c(\boldsymbol{z}) = S_2^c(\boldsymbol{z})$, which would result in (8) (and the well-known product-form solution indeed satisfies it).

Remark 9 Consider an even simpler system with internal transitions: a single service facility, Poisson(λ) arrivals, feedback after service with probability p, system departure with probability $1-p$. Now the balance equations become, with X denoting number of customers and F denoting that we view the system just after a departure from the queue and just before its return to the queue:

$$\begin{aligned} & \lambda \mathbb{P}(X = j|A) + \lambda \mathbb{P}(X = j-1|D) + \frac{1-p}{p} \lambda \mathbb{P}(X = j-1|F) \\ = \quad & \lambda \mathbb{P}(X = j-1|A) + \lambda \mathbb{P}(X = j|D) + \frac{1-p}{p} \lambda \mathbb{P}(X = j-1|F). \end{aligned} \tag{18}$$

It results in cancellation of the two feedback terms, and in the statement that in this system (as we already know, for any service time distribution) the distributions at arrival epochs and at real departure epochs are equal.

Let us finally consider a network of m service facilities, with independent external Poisson arrival processes, and again with continuous service time distributions. We have Markovian routing, a customer moving from Q_i to Q_k with probability p_{ik} and leaving the system after its service completion in Q_i with probability p_{i0}, $i, k \in J$. Define Λ_i as the total flow through Q_i per time unit, $i \in J$; these Λ_i are the unique solution of the set of equations

$$\Lambda_i = \lambda_i + \sum_{k=1}^{m} \Lambda_k p_{ki}, \quad i \in J. \tag{19}$$

The balance equations become, with I_{ik} denoting that we view the system just after a departure from Q_i and just before the arrival of the departing customer at Q_k: for $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ and $\boldsymbol{j} = (j_1, j_2, \ldots, j_m)$,

$$\begin{aligned} & \sum_{i=1}^{m} \lambda_i \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|A_i) + \sum_{i=1}^{m} \Lambda_i p_{i0} \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_i|D_i) + \sum_{i=1}^{m} \sum_{k=1}^{m} \Lambda_i p_{ik} \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_i|I_{ik}) \\ = & \sum_{i=1}^{m} \Lambda_i p_{i0} \mathbb{P}(\boldsymbol{X} = \boldsymbol{j}|D_i) + \sum_{i=1}^{m} \lambda_i \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_i|A_i) + \sum_{i=1}^{m} \sum_{k=1}^{m} \Lambda_i p_{ik} \mathbb{P}(\boldsymbol{X} = \boldsymbol{j} - \boldsymbol{e}_k|I_{ik}). \end{aligned} \tag{20}$$

It is obviously seen that the (PGF of the) probabilities, given that we observe just after a real departure from Q_i or that we observe just after a departure from Q_i that will in an instant result in an arrival at Q_k, are the same. Such results can be found in [4, 11]. We will formally verify (20) as a special case of a more general network (see the end of Section 6).

In Example 3 of the next section, we use (20) to provide an alternative proof for the joint queue-length distribution in a queueing network with a single roving server as studied in [4, 11].

5 Examples

In this section we consider three multiclass queues to which our balance results apply. In the first example, a non-preemptive priority queue with P customer classes, we first verify the equality between the PGF's as given by Theorem 1 for the case $P = 2$, and subsequently point out how one may use the theorem to obtain the steady-state joint queue-length distribution in that example for a P-class queue. In the second example, a model with two customer classes and priority for the longer queue, application of Theorem 1 immediately yields the PGF of the steady-state joint queue-length distribution at an arbitrary epoch. In the third example, we show that the joint queue-length distribution in a queueing network with a roving server, as derived in [11], could also have been derived from the balance equation (20).

5.1 Example 1: Non-preemptive priority queues

Consider the $M/G/1$ queue with P classes of customers, with nonpreemptive priority in descending order $1, 2, \ldots, P$ (so Class 1 has the highest priority). Takagi ([12], Formula (2.87) on p. 311) presents the PGF $\Pi(z_1, z_2, \ldots, z_P)$ of the steady-state joint queue-length distribution immediately after an arbitrary customer departure epoch. For $P = 2$ he also obtains the PGF $P(z_1, z_2, \ldots, z_P)$ of the steady-state joint queue-length distribution at an arbitrary epoch ([12], Formula (5.82b) on p. 397). We have verified that, indeed, for $P = 2$ classes one has,

$$\lambda_1(1 - z_1)S_{c_1}(z_1, z_2) + \lambda_2(1 - z_2)S_{c_2}(z_1, z_2) = (\lambda_1(1 - z_1) + \lambda_2(1 - z_2))P(z_1, z_2).$$

Starting-point for this verification was the obvious set of relations, with $\beta_i(z_1, z_2)$ the PGF of the numbers of arrivals at both queues during one service of a class-i customer, $i = 1, 2$:

$$\Pi_1(z_1, z_2) := \frac{\lambda_1}{\lambda} S_{c_1}(z_1, z_2) = \frac{\Pi(z_1, z_2) - \Pi(0, z_2)}{z_1}\beta_1(z_1, z_2) + \Pi(0, 0)\frac{\lambda_1}{\lambda}\beta_1(z_1, z_2), \qquad (21)$$

$$\Pi_2(z_1, z_2) := \frac{\lambda_2}{\lambda} S_{c_2}(z_1, z_2) = \frac{\Pi(0, z_2) - \Pi(0, 0)}{z_2}\beta_2(z_1, z_2) + \Pi(0, 0)\frac{\lambda_2}{\lambda}\beta_2(z_1, z_2). \qquad (22)$$

Here $\Pi_i(z_1, z_2)$ is the PGF of the steady-state joint queue-length distribution immediately after the departure of a class-i customer, with indicator function $1(\text{departing customer is of class } i)$, $i = 1, 2$. The factors $\frac{\lambda_i}{\lambda}$ in the lefthand side of (21) and (22) are needed because the $S_{c_i}(z_1, z_2)$ are conditional PGF's, the condition being that the departing customer is of class i.

This example clearly demonstrates the value of our general balance equations. Besides providing a much shorter proof for Takagi's Formula (5.82b), it also allows us to extend his result to the case of $P(> 2)$ customer classes, by using the expressions for $\frac{\lambda_i}{\lambda} S_{c_i}(z_1, z_2, \ldots, z_P)$ that follow from Takagi's Formula (2.87) for $\Pi(z_1, z_2, \ldots, z_P)$.

5.2 Example 2: Priority for the longer queue

Consider a model of one server and two queues. Each queue has its own Poisson arrival process and service time distribution. After a service completion, the server proceeds with a customer from the longest queue, if the queue lengths are unequal; if the queue lengths are equal, the server chooses a customer from queue Q_i with probability α_i, $i = 1, 2$. Cohen [6] has derived the PGF $\Pi(z_1, z_2) = \mathbb{E}[z_1^{X_1} z_2^{X_2}]$ of the steady-state joint queue-length distribution immediately after an arbitrary customer departure epoch, by solving a Riemann-type boundary value problem. In the process, he also obtained the following PGFs, that naturally arise in this *Priority for the longer queue* model: $\mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 > X_2\}}]$, $\mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 < X_2\}}]$, and $\mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 = X_2 > 0\}}]$. Below we first show how one can obtain the PGFs $\Pi_i(z_1, z_2)$ of the steady-state joint queue-length distribution immediately after the departure of a customer from Q_i, $i = 1, 2$ (we stick as much as possible to the notation of Example 1). By considering the joint queue-length distribution at two consecutive departure epochs, and with $\beta_i(z_1, z_2)$ denoting the PGF of the numbers of arrivals at both queues during one service of a customer from Q_i, we can write:

$$\Pi_1(z_1, z_2) = \mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 > X_2\}}] \frac{\beta_1(z_1, z_2)}{z_1} + \alpha_1 \mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 = X_2 > 0\}}] \frac{\beta_1(z_1, z_2)}{z_1} + \mathbb{P}(X_1 = X_2 = 0) \frac{\lambda_1}{\lambda_1 + \lambda_2} \beta_1(z_1, z_2), \quad (23)$$

$$\Pi_2(z_1, z_2) = \mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 < X_2\}}] \frac{\beta_2(z_1, z_2)}{z_2} + \alpha_2 \mathbb{E}[z_1^{X_1} z_2^{X_2} 1_{\{X_1 = X_2 > 0\}}] \frac{\beta_2(z_1, z_2)}{z_2} + \mathbb{P}(X_1 = X_2 = 0) \frac{\lambda_2}{\lambda_1 + \lambda_2} \beta_2(z_1, z_2). \quad (24)$$

The queue-length PGFs in the two righthand sides are derived by Cohen [6], and thus we obtain $\Pi_i(z_1, z_2)$, $i = 1, 2$. This immediately leads to $S_{c_i}(z_1, z_2)$, $i = 1, 2$, as in Example 1. Subsequently, Theorem 1 gives the PGF of the steady-state joint queue-length distribution at an arbitrary epoch. It should be noticed that it is not at all easy to obtain this PGF in another way, for this non-Markovian model; the *Priority for the longer queue* model is a difficult queueing model. In the case of exponential service time distributions, with equal arrival and service rates at the two queues and $\alpha_1 = \alpha_2$, Zheng and Zipkin [13] present a recursive method to obtain this PGF, while Flatto [8] for this case (but allowing preemption) obtains the queue-length PGF by solving a boundary value problem.

5.3 Example 3: A queueing network with a roving server

We consider a network of m queues with Markovian customer routing, as described in the final paragraphs of the previous section. In this particular example, we assume that a *single* server visits the queues in a fixed, cyclic order, requiring a switch-over time S_i to move from Q_i to Q_{i+1}. We do not make any assumptions regarding the service disciplines at each queue. This model, which can be regarded as a polling model with customer routing, has been studied by Sidi, Levy and Fuhrmann [11] who refer to this model as a queueing network with a roving server. Sidi et al. obtain the joint queue-length distribution at arbitrary moments, as well as the joint queue-length distribution at departure epochs. The waiting-time distributions are obtained in a different paper [4]. For us, it is slightly more convenient to refer to this latter paper in the analysis below, because the authors in [4] use the same definition of

$V_i^c(\boldsymbol{z})$, the PGF of the joint queue length at departure epochs, just after a departure from Q_i and just *before* the arrival of the departing customer at the next queue.

Take the formulas (3.2)–(3.6) of [4]. From (3.2), which is the counterpart of our (2), one can express (in the notation of the present paper) the differences of PGFs at visit beginning and visit completion epochs into those at service beginning and service completion epochs:

$$\frac{V_i^b(\boldsymbol{z}) - V_i^c(\boldsymbol{z})}{\Lambda_i \mathbb{E}C} = S_i^b(\boldsymbol{z}) - S_i^c(\boldsymbol{z}) P_i(\boldsymbol{z}), \quad i = 1, 2, \ldots, m. \tag{25}$$

Here $P_i(\boldsymbol{z}) := p_{i0} + \sum_{k=1}^m p_{ik} z_k$, and $\mathbb{E}C = s/(1-\rho)$ with $\rho := \sum_{i=1}^m \Lambda_i b_i$. Next use our relation (3) to express $S_i^b(\boldsymbol{z})$ into $S_i^c(\boldsymbol{z})$. Subsequently express $L(\boldsymbol{z})$, in (3.4) of [4], which is the counterpart of (1) above, in differences $V_i^b(\boldsymbol{z}) - V_i^c(\boldsymbol{z})$, as was also done in [3]. This gives

$$\sum_{i=1}^m \lambda_i (1 - z_i) L(\boldsymbol{z}) = \sum_{i=1}^m \Lambda_i (P_i(\boldsymbol{z}) - z_i) S_i^c(\boldsymbol{z}). \tag{26}$$

This is indeed in agreement with (20): the LHS of (26) gives the first and the fifth term in (20). The last term in the RHS gives the second plus the third term in (20), once we realize that $p_{i0} + \sum_{k=1}^m p_{ik} = 1$, and that the conditional probabilities both refer to a service completion in Q_i, no matter whether the condition is D_i or I_{ik}. The first term in the RHS gives the fourth plus fifth term in (20). One could argue that some results in [4] and [11] could have been derived faster by starting from (26).

6 Formal derivations under a general framework

So far, we have derived distributional relationships at arrival and departure instants for various queues and their network models using flow balance equations. In this section, we aim to derive them in a unified way under general settings. Roughly speaking, these settings allow simultaneous external arrivals, simultaneous departures and routing at different stations; however, we do not allow that an external arrival coincides with a departure. We use their time evolutions in sample paths for deriving the relationships rather than using flow balance.

We describe a queueing network system under a fairly general framework. We consider an open queueing network system with m queues, where queues uniquely belong to service facilities, which are called stations. Queues in the same station may be distinguished by customer classes. Each station may have multiple servers, which may change in time. External arrivals at queues are general as long as they satisfy certain stationarity conditions. Customers completing service may be routed among queues depending on the state of the whole system. Thus, this model is quite general, very flexible and covers all examples in the previous section.

To describe this model, we introduce a stochastic process. Queues are still indexed by $J = \{1, 2, \ldots, m\}$. Let

$$\boldsymbol{X}(t) = (X_1(t), \ldots, X_m(t)),$$

where $X_i(t)$ represents the length of queue i at time t, which includes customers in service. Here, each queue belongs to a single station. There is a mapping from queues to stations, which will be given when needed.

In addition to $\boldsymbol{X}(t)$, the following counting processes count the number of specified events until time $t \geq 0$ for $i \in J$,

- $N_i^e(t)$ - external arrivals at queue i,
- $N_i^d(t)$ - departures from queue i,
- $N_i^r(t)$ - internal arrivals at queue i (transition from some queue).

With $\boldsymbol{N}^u(t) = (N_1^u(t), \ldots, N_m^u(t))$ for $u = e, d, r$, we consider the process

$$\boldsymbol{Z}(t) \equiv (\boldsymbol{X}(t), \boldsymbol{N}^e(t), \boldsymbol{N}^d(t), \boldsymbol{N}^r(t)) \, .$$

All processes are assumed right-continuous with left limits. Let $\Delta \boldsymbol{X}(t) = \boldsymbol{X}(t) - \boldsymbol{X}(t-)$. $\Delta \boldsymbol{N}^u(t)$ is similarly defined and is in $\mathbb{Z}_+^m$ for $u = e, d, r$, where $\mathbb{Z}_+$ is the set of nonnegative integers.

For $u = e, d, r$, denote

$$|\boldsymbol{N}^u|(t) = \sum_{i \in J} N_i^u(t)$$

and assume that

i) $\boldsymbol{X}(0), \boldsymbol{N}^e(t), \boldsymbol{N}^d(t), \boldsymbol{N}^r(t)$ are all finite (in $\mathbb{Z}_+^m$) for each $t \geq 0$.

ii) $\Delta|\boldsymbol{N}^e|(t)\Delta|\boldsymbol{N}^d|(t) = 0$ for each $t \geq 0$. That is, external arrivals and service completions can not occur simultaneously.

We also need to define the intermediate state

$$\boldsymbol{X}^d(t) = \boldsymbol{X}(t) - \Delta \boldsymbol{N}^r(t) \in \mathbb{Z}_+^m \, . \tag{27}$$

It differs from $\boldsymbol{X}(t)$ only at departure epochs and it describes the state "after" a departure and "before" an arrival at a different queue.

Clearly, the following dynamics holds.

$$\boldsymbol{X}(t) = \boldsymbol{X}(0) + \boldsymbol{N}^e(t) - \boldsymbol{N}^d(t) + \boldsymbol{N}^r(t) \in \mathbb{Z}_+^m \, . \tag{28}$$

Because of i), $\boldsymbol{X}(t)$ and $\boldsymbol{X}^d(t)$ are also finite. It may be natural to assume that $|\boldsymbol{N}^r|(t) \leq |\boldsymbol{N}^d|(t)$ for $t \geq 0$, but we do not require it in this section.

Thus, $\boldsymbol{X}(t)$ is the state of the system at time t of an input-output system driven by counting processes $\boldsymbol{N}^e, \boldsymbol{N}^d, \boldsymbol{N}^r$. The dynamics of (27) and (28) indicates that we adopt the *departure first* framework. We have used queueing terminologies, but our results are valid as long as the above mathematical assumptions and (28) are satisfied.

In general, $|\boldsymbol{N}^e|(t), |\boldsymbol{N}^d|(t)$ and $N_i^e(t)$ and $N_i^d(t)$ may have jumps greater than one, which is not convenient to describe the time evolution of $\boldsymbol{Z}(t)$. Thus, for $u = e, d$, we introduce

$$|\tilde{\boldsymbol{N}}^u|(t) = \sum_{0 < s \leq t} 1(\Delta|\boldsymbol{N}^u|(s) \geq 1), \qquad \tilde{N}_i^u(t) = \sum_{0 < s \leq t} 1(\Delta N_i^u(s) \geq 1),$$

then $\Delta|\tilde{\boldsymbol{N}}^u|(t) \leq 1$ and $\Delta \tilde{N}_i^u(t) \leq 1$, that is, $|\tilde{\boldsymbol{N}}^u|$ and $\tilde{N}_i^u$ are simple point processes for $u = e, d$. Set $t_0^e = t_0^d = t_{i,0}^e = t_{i,0}^d = 0$ for $i \in J$, and for $n \geq 1$ and $i \in J$ let $t_n^e, t_n^d, t_{i,n}^e, t_{i,n}^d$ be the n^{th} jump epoch of $|\tilde{\boldsymbol{N}}^e|, |\tilde{\boldsymbol{N}}^d|, \tilde{N}_i^e, \tilde{N}_i^d$, respectively (of course, if the corresponding process is not terminating and such an epoch exists).

Another basic assumption on the counting processes is

iii) There exist finite and positive numbers λ^u, $u = e, d$ such that

$$\lambda^u = \lim_{t\to\infty} \frac{1}{t}|\tilde{\boldsymbol{N}}^u|(t)\ , \tag{29}$$

a.s. (almost surely) w.r.t. the underlying probability measure $\mathbb{P}$.

We further assume the following ergodic type conditions.

iv) There exist probability distributions π^e and π^d such that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) = \pi^e(\boldsymbol{x},\boldsymbol{y}), \quad a.s., \quad \boldsymbol{x},\boldsymbol{y} \in \mathbb{Z}_+^m, \tag{30}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} 1(\boldsymbol{X}^d(t_\ell^d) = \boldsymbol{x}, \Delta\boldsymbol{N}^d(t_\ell^d) = \boldsymbol{y}, \Delta\boldsymbol{N}^r(t_\ell^d) = \boldsymbol{z}) = \pi^d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}), \quad a.s.,$$

$$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathbb{Z}_+^m. \tag{31}$$

From the definitions in iv), π^e and π^d are considered as the embedded stationary distributions just before arrival epochs and just after departure epochs but before internal arrivals, respectively. They correspond to Palm distributions concerning their counting processes in the time stationary framework (e.g., see [1]).

Since the process $\boldsymbol{X}(t)$ is vector valued, it is not so convenient for manipulations. So, we introduce a test function $f : \mathbb{Z}_+^m \to \mathbb{R}$. Under the setting i)–iv), we will derive distributional relationships among characteristics at different embedded instants using the test function f. For this, we need the following lemma.

Lemma 1 *If (30) holds, then, for any bounded function $g : \mathbb{Z}_+^{2m} \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} g(\boldsymbol{X}(t_\ell^e-), \Delta\boldsymbol{N}^e(t_\ell^e)) = \sum_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}_+^m} g(\boldsymbol{x},\boldsymbol{y})\pi^e(\boldsymbol{x},\boldsymbol{y}), \quad \text{a.s.} \tag{32}$$

Similarly, if (31) holds, then, for any bounded function $h : \mathbb{Z}_+^{3m} \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} h(\boldsymbol{X}^d(t_\ell^d), \Delta\boldsymbol{N}^d(t_\ell^d), \Delta\boldsymbol{N}^r(t_\ell^d)) = \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{Z}_+^m} h(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\pi^d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}), \quad \text{a.s.} \tag{33}$$

This lemma may look obvious, but its proof is not immediate because we need to verify the exchange of limits. So, we prove it in Appendix A.

We are now ready to prove distributional relationships. First, we denote the expectations under π^e and π^d by $\mathbb{E}^e$ and $\mathbb{E}^d$, respectively. That is,

$$\mathbb{E}^e g(\boldsymbol{X},\boldsymbol{Y}) = \sum_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{Z}_+^m} g(\boldsymbol{x},\boldsymbol{y})\pi^e(\boldsymbol{x},\boldsymbol{y}), \tag{34}$$

$$\mathbb{E}^d h(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{Z}_+^m} h(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\pi^d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}). \tag{35}$$

Note that $\boldsymbol{Y}$ in $\mathbb{E}^e$ represents sizes of externally arriving batches, while $\boldsymbol{Y}$ in $\mathbb{E}^d$ represents sizes of departing batches.

Theorem 3 *Under the setting i)–iv), for any bounded function $f : \mathbb{Z}_+^m \to \mathbb{R}$, we have*

$$\lambda^e \mathbb{E}^e\big[f(\boldsymbol{X}+\boldsymbol{Y}) - f(\boldsymbol{X})\big] + \lambda^d \mathbb{E}^d\big[f(\boldsymbol{X}+\boldsymbol{Z}) - f(\boldsymbol{X})\big] = \lambda^d \mathbb{E}^d\big[f(\boldsymbol{X}+\boldsymbol{Y}) - f(\boldsymbol{X})\big]. \tag{36}$$

Proof Since $f(\boldsymbol{X}(t))$ changes in time only at the counting instants t_n^e or t_n^d, we have

$$f(\boldsymbol{X}(t)) - f(\boldsymbol{X}(0)) = \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^e|(t)} \Delta f(\boldsymbol{X}(t_\ell^e)) + \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} \Delta f(\boldsymbol{X}(t_\ell^d)). \tag{37}$$

Recalling (27), we decompose the last sum as

$$\begin{aligned}\sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} \Delta f(\boldsymbol{X}(t_\ell^d)) &= \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}(t_\ell^d)) - f(\boldsymbol{X}^d(t_\ell^d))) + \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}^d(t_\ell^d)) - f(\boldsymbol{X}(t_\ell^d-)))\\ &= \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}^d(t_\ell^d) + \Delta \boldsymbol{N}^r(t_\ell^d)) - f(\boldsymbol{X}^d(t_\ell^d)))\\ &\quad + \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}^d(t_\ell^d)) - f(\boldsymbol{X}^d(t_\ell^d) + \Delta \boldsymbol{N}^d(t_\ell^d))),\end{aligned} \tag{38}$$

using the fact that $\boldsymbol{X}(t_\ell^d) - \boldsymbol{X}(t_\ell^d-) = -\Delta \boldsymbol{N}^d(t_\ell^d) + \Delta \boldsymbol{N}^r(t_\ell^d)$ by (28), equivalently, $\boldsymbol{X}(t_\ell^d-) = \boldsymbol{X}^d(t_\ell^d) + \Delta \boldsymbol{N}^d(t_\ell^d)$ by (27). It follows from (37) and (38) that

$$\begin{aligned}&\sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^e|(t)} (f(\boldsymbol{X}(t_\ell^e-) + \Delta \boldsymbol{N}^e(0)) - f(\boldsymbol{X}(t_\ell^e-))) + \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}^d(t_\ell^d) + \Delta \boldsymbol{N}^r(t_\ell^d)) - f(\boldsymbol{X}^d(t_\ell^d)))\\ &\quad = \sum_{\ell=1}^{|\tilde{\boldsymbol{N}}^d|(t)} (f(\boldsymbol{X}^d(t_\ell^d) + \Delta \boldsymbol{N}^d(t_\ell^d)) - f(\boldsymbol{X}^d(t_\ell^d))) + f(\boldsymbol{X}(t)) - f(\boldsymbol{X}(0)).\end{aligned} \tag{39}$$

Dividing both sides of this equation by t and letting $t \to \infty$ yields (36) by (29)–(31) and Lemma 1 because f is bounded. □

The assumptions of Theorem 3 allow simultaneous external arrivals and simultaneous departure and routing. They exclude that arrivals and departures occur simultaneously, but may be too general for queueing networks. To make them more specific, we make the following assumption.

v) There exist finite and positive numbers λ_A^d for nonempty $A \subset J$, that is, $A \in 2^J \setminus \{\emptyset\}$, such that

$$\lambda_A^d = \lim_{t\to\infty} \frac{1}{t} \tilde{N}_A^d(t), \qquad a.s., \tag{40}$$

where, with notation $S_A \equiv \{\boldsymbol{x} \in \mathbb{Z}_+^m; x_i > 0, i \in A, x_j = 0, j \in J \setminus A\}$,

$$\tilde{N}_A^d(t) = \sum_{0<s\le t} 1(\Delta \boldsymbol{N}^d(s) \in S_A). \tag{41}$$

Note that $\tilde{N}_A^d$ counts instants when departures occur simultaneously from queues $i \in A$ but there is no departure from queue $j \in J \setminus A$, while $\Delta\tilde{N}_A^d(t)\Delta\tilde{N}_B^d(t) = 0$ if $A \neq B$. Thus, the setting i)–v) still allows batch arrivals and batch departures and simultaneous transfer of customers in a departing batch.

We will use the following notation. For each $A \in 2^J \setminus \{\emptyset\}$, let

$$\pi_A^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \begin{cases} \frac{\lambda^d}{\lambda_A^d}\pi^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}), & \boldsymbol{y} \in S_A, \\ 0, & \boldsymbol{y} \in \mathbb{Z}_+^m \setminus S_A. \end{cases}$$

Since $\tilde{N}_A^d$ exclusively counts the increasing epochs of $|\tilde{\boldsymbol{N}}^d|$ for different A's, we have

$$|\tilde{\boldsymbol{N}}^d|(t) = \sum_{A \in 2^J \setminus \{\emptyset\}} \tilde{N}_A^d(t), \tag{42}$$

which implies that $\lambda = \sum_{A \in S_A \setminus \{\emptyset\}} \lambda_A$, and therefore π_A^d is a probability distribution on $\mathbb{Z}_+^{3m}$, which can be restricted to $\mathbb{Z}_+^m \times S_A \times \mathbb{Z}_+^m$.

Let $t_{A,n}^d$ be the n^{th} jump epoch of $\tilde{N}_A^d$. Just as Lemma 1, the following lemma plays a key role; it is proved in Appendix B.

Lemma 2 *Under the setting i)–v), there exist probability distributions π_i^d such that, for any bounded function $h : \mathbb{Z}_+^{3m} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} h(\boldsymbol{X}^d(t_{A,\ell}^d), \Delta\boldsymbol{N}^d(t_{A,\ell}^d), \Delta\boldsymbol{N}^r(t_{A,\ell}^d)) = \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathbb{Z}_+^m} h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\pi_A^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}), \text{ a.s.}, \qquad i \in J. \tag{43}$$

By (42), Theorem 3 and Lemma 2 yield the following corollary. As with $\mathbb{E}^e$ and $\mathbb{E}^d$, $\mathbb{E}_A^d$ stands for the expectation under π_A^d.

Corollary 1 *Under the setting i)–v), for any bounded function $f : \mathbb{Z}_+^m \to \mathbb{R}$,*

$$\begin{aligned} \lambda^e \mathbb{E}^e \left[f(\boldsymbol{X} + \boldsymbol{Y}) - f(\boldsymbol{X}) \right] + \sum_{A \in 2^J \setminus \{\emptyset\}} \lambda_A^d \mathbb{E}_A^d \left[f(\boldsymbol{X} + \boldsymbol{Z}) - f(\boldsymbol{X}) \right] \\ = \sum_{A \in 2^J \setminus \{\emptyset\}} \lambda_A^d \mathbb{E}_A^d \left[f(\boldsymbol{X} + \boldsymbol{Y}) - f(\boldsymbol{X}) \right]. \end{aligned} \tag{44}$$

Remark 10 *If $\Delta\tilde{N}_i^d(t_{j,n}^d) = 0$ for all $i \neq j$, then $\lambda_A^d > 0$ only if A is a singleton. In this case, the summations over A in (44) can be reduced to those over $i \in J$, replacing A by i.*

Until now, our distributional relationship may still be too general because no assumption is made on how the counting processes are generated from $\boldsymbol{X}(t)$ and other information. To describe this, a filtration is convenient. Let $\mathcal{F}_t$ be the σ-field generated by all events up to time t, and let $\mathcal{F}_{t-} = \sigma(\cup_{u<t}\mathcal{F}_u)$, that is, $\mathcal{F}_{t-}$ is a σ-field generated by all events before time t. For a stopping time τ, let $\mathcal{F}_{\tau-} = \sigma(\mathcal{F}_0, \{A \cap \{t < \tau\} \in \mathcal{F}_t\})$, where $\sigma(\mathcal{A})$ is the σ-field generated by a family of events $\mathcal{A}$. Using the filtration, the following assumptions are typically used under the setting i)–v).

(a1) $t_n^e, t_{i,n}^d$ are stopping times with respect to $\{\mathcal{F}_t; t \geq 0\}$. This can always be realized by choosing a sufficiently large $\mathcal{F}_t$.

(a2) $\Delta \boldsymbol{N}^e(t_n^e)$ is independent of $\mathcal{F}_{t_n^e-}$. This is, sizes of batch arrivals are independent of the state of the system just before their arrival epochs.

(a3) $\Delta|\boldsymbol{N}^d|(t_n^d) = 1$. That is, departures singly occur from one queue at a time.

(a4) $\Delta N_j^r(t_{i,n}^d) \leq 1$ for $j \in J$, and $\Delta \boldsymbol{N}^r(t_{i,n}^d)$ is in the σ-field generated by $\mathcal{F}_{t_{i,n}^d-}$ and $\Delta \boldsymbol{N}^d(t_{i,n}^d)$.

By (a3), $\tilde{N}_A^d(t) \equiv 0$ if A is not a singleton. Thus, we write $\tilde{N}_A^d(t)$ as $\tilde{N}_i^d(t)$ for $A = \{i\}$. Similarly, π_A^d is written as π_i^d for $A = \{i\}$. Under the setting i)–v) and the assumptions (a1)–(a4), $\Delta N_j^r(t_{i,\ell}^d) \leq 1$, and therefore Lemma 2 yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} 1(\boldsymbol{X}^d(t_{i,\ell}^d) = \boldsymbol{x}, \Delta N_i^d(t_{i,\ell}^d) = 1, \Delta N_j^r(t_{i,\ell}^d) = 1) = \pi_i^d(\boldsymbol{x}, \boldsymbol{e}_i, \boldsymbol{e}_j),$$

which is denoted by $\pi_{ij}^d(\boldsymbol{x})$. We here recall that $\boldsymbol{e}_i \in \mathbb{Z}_+^m$ is the unit vector whose i-th entry is one and the other entries are zero. Thus, applying Corollary 1 for $f(\boldsymbol{x}) = \boldsymbol{z}^{\boldsymbol{x}}$, where we recall that $\boldsymbol{z}^{\boldsymbol{x}} = \prod_{i\in J} z_i^{x_i}$, we have the following relationship.

Corollary 2 *Under the settings i)–v) and assumptions (a1)–(a4), for $\boldsymbol{z} = (z_1, \ldots, z_m)$ satisfying $|z_i| \leq 1$ for $i \in J$,*

$$\lambda^e\big(1 - \mathbb{E}^e[\boldsymbol{z}^{\boldsymbol{Y}}]\big)\varphi^e(\boldsymbol{z}) + \sum_{j\in J}(1 - z_j) \sum_{i\in J} \lambda_i^d \varphi_{ij}^d(\boldsymbol{z}) = \sum_{i\in J} \lambda_i^d (1 - z_i)\varphi_i^d(\boldsymbol{z}), \tag{45}$$

where

$$\varphi^e(\boldsymbol{z}) = \mathbb{E}^e[\boldsymbol{z}^{\boldsymbol{X}}], \qquad \varphi_i^d(\boldsymbol{z}) = \mathbb{E}_i^d[\boldsymbol{z}^{\boldsymbol{X}}], \qquad \varphi_{i,j}^d(\boldsymbol{z}) = \sum_{\boldsymbol{x}\in\mathbb{Z}_+^m} \boldsymbol{z}^{\boldsymbol{x}} \pi_{ij}^d(\boldsymbol{x}), \qquad i, j \in J.$$

Remark 11 Under the assumptions of this corollary, the routing of departing customers may depend on all queue lengths in the network.

Corollary 2 is specialized to Corollary 3 if external arrivals to queues occur one at a time. Namely,

vi) No simultaneous arrivals occur, and there exist finite numbers (some, but not all, possibly zero) λ_k^e for $k \in J$ such that

$$\lambda_k^e = \lim_{t\to\infty} \frac{1}{t} \tilde{N}_k^e(t), \quad a.s., \qquad k \in J. \tag{46}$$

Corollary 3 *Under the assumptions of Corollary 2, if the assumption vi) holds, then*

$$\pi_k^e(\boldsymbol{x}, y_k) \equiv \frac{\lambda^e}{\lambda_k^e} \pi^e(\boldsymbol{x}, y_k), \quad k \in \{i|\lambda_i^e > 0\},$$

and is defined arbitrarily when $\lambda_k^e = 0$, is a probability distribution on $\mathbb{Z}_+^{m+1}$, and (45) becomes

$$\sum_{k\in J} \lambda_k^e \big(1 - \mathbb{E}^e\big[z_k^{Y_k}\big]\big)\varphi_k^e(\boldsymbol{z}) + \sum_{j\in J}(1 - z_j) \sum_{i\in J} \lambda_i^d \varphi_{ij}^d(\boldsymbol{z}) = \sum_{i\in J} \lambda_i^d (1 - z_i)\varphi_i^d(\boldsymbol{z}), \tag{47}$$

where φ_k^e is the generating function of $\boldsymbol{X}$ under the conditional distribution π_k^e.

Corollary 4 *Under the assumptions of Corollary 3, if the event* $\{\Delta N_j^r(t_{i,\ell}^d) = 1\}$ *is independent of* $\mathcal{F}_{t_{i,\ell}^d-}$*, then there exist* $p_{ij} \geq 0$ *such that* $\pi_i^d(\boldsymbol{x}, 1, \boldsymbol{e}_j) = \pi_i^d(\boldsymbol{x}, 1)p_{ij}$*, and (47) becomes*

$$\sum_{k\in J_e} \lambda_k^e \big(1 - \mathbb{E}^e\big[z_k^{Y_k}\big]\big)\varphi_k^e(\boldsymbol{z}) + \sum_{i\in J} \lambda_i^d \varphi_i^d(\boldsymbol{z}) \sum_{j\in J} p_{ij}(1 - z_j) = \sum_{i\in J} \lambda_i^d (1 - z_i)\varphi_i^d(\boldsymbol{z}). \tag{48}$$

We now list how major results in Sections 2–5, including Theorems 1 and 2, are obtained as special cases of the general results in this section. We first note that, if nonzero N_k^e for $k \in J$ are independent Poisson processes, then by PASTA the embedded stationary distributions π^e and π_k^e are identical with the time stationary distributions.

1) Theorem 1 is a special case of (45) of Corollary 2 in which there is no routing.

2) In view of Remark 10, Theorem 2 is a special case of (44) of Corollary 1 in which there is no routing, the external arrival batch $\boldsymbol{Y}(t_n^e)$ is independent of $\mathcal{F}_{t_n^e-}$ and the departing batch size Y_i from queue i is some constant K_i. In this case, it is easy to see that $\lambda_i^d \mathbb{E}[Y_i] = \lambda^e / K_i$, and we obtain (14) from (44).

3) (17) and (20) in Section 4 (and therefore (26) in Section 5.3) may look different from (48), but they can be considered as its special cases if we rightly replace the conditional probabilities in (20).

Remark 12 Notice that our setup includes the finite buffer case. This is done by having no arrivals to a queue during times in which it is saturated. This type of dependence is allowed by our setup.

7 Distributional relationship up to a given time

The purpose of this section is to derive a non-stationary version of Theorem 3, a distributional relationship up to a given time. We adopt the settings i)–iv) of Section 6, and consider the process $\boldsymbol{Z}(t)$ introduced in the beginning of that section. We first define the expected relative frequencies for bounded test functions g, h from $\mathbb{Z}_+^{2m}, \mathbb{Z}_+^{3m}$ to $\mathbb{R}$ up to time t as

$$R_t^e g = \frac{1}{|\tilde{\boldsymbol{N}}^e|(t)} \sum_{n=1}^{|\tilde{\boldsymbol{N}}^e|(t)} g(\boldsymbol{X}(t_n^e-), \Delta\boldsymbol{N}^e(t_n^e)),$$
$$R_t^d h = \frac{1}{|\tilde{\boldsymbol{N}}^d|(t)} \sum_{n=1}^{|\tilde{\boldsymbol{N}}^e|(t)} h(\boldsymbol{X}^d(t_n^d), \Delta\boldsymbol{N}^d(t_n^d), \Delta\boldsymbol{N}^r(t_n^d)).$$

For each bounded function $f : \mathbb{Z}_+^m \to \mathbb{R}$, we define the following test functions.

$$g^e(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}), \qquad g_+^e(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x} + \boldsymbol{y}),$$
$$h^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x}), \qquad h_-^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x} + \boldsymbol{y}), \qquad h_+^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x} + \boldsymbol{z}).$$

Let

$$\lambda^e(t) = \frac{1}{t}|\tilde{\boldsymbol{N}}^e|(t), \qquad \lambda^d(t) = \frac{1}{t}|\tilde{\boldsymbol{N}}^d|(t).$$

Then, (39) yields the following lemma.

Lemma 3 *Under the setting i)–v), for any bounded function $f : \mathbb{Z}_+^m \to \mathbb{R}$, we have, for any $t > 0$,*

$$\lambda^e(t)\big(R_t^e g_+^e - R_t^e g^e\big) + \lambda^d(t)\big(R_t^d h_+^d - R_t^d h^d\big) - \lambda^d(t)\big(R_t^d h_-^d - R_t^d h^d\big) = \frac{1}{t}\big(f(\boldsymbol{X}(t)) - f(\boldsymbol{X}(0))\big). \tag{49}$$

We may interpret Lemma 3 as a transient version of Theorem 3. It is notable that (49) holds without any stability condition, and its right-hand side vanishes as $t \to \infty$ at most in linear order of t^{-1} because f is bounded. If there exists a unique probability measure such that $(\boldsymbol{X}(t), \Delta\boldsymbol{N}^e(t), \Delta\boldsymbol{N}^d(t), \Delta\boldsymbol{N}^r(t))$ is stationary, then $R_t^e g, R_t^d h$ converge to the corresponding expectations under the Palm distributions involving $|\tilde{\boldsymbol{N}}^e|, |\tilde{\boldsymbol{N}}^d|$, respectively. Thus, we have

$$\begin{aligned}
&\lim_{t\to\infty} R_t^e g^e = \mathbb{E}^e f(\boldsymbol{X}), \qquad \lim_{t\to\infty} R_t^e g_+^e = \mathbb{E}^e f(\boldsymbol{X}+\boldsymbol{Y}),\\
&\lim_{t\to\infty} R_t^d h^d(\boldsymbol{x}) = \mathbb{E}^d f(\boldsymbol{X}), \qquad \lim_{t\to\infty} R_t^d h_-^d = \mathbb{E}^e f(\boldsymbol{X}+\boldsymbol{Y}),\\
&\lim_{t\to\infty} R_t^d h_+^d = \mathbb{E}^d f(\boldsymbol{X}+\boldsymbol{Z}),
\end{aligned}$$

and we recover (36) from (49). Corollary 1, (45) and (47) are similarly obtained. We omit the routine details.

8 Concluding remarks

This paper derives a distributional relationship, at different embedded epochs, for analyzing queues and their networks. As shown in Section 6, it has different forms according to the abstraction level of the model. In Sections 2–5 we also discussed how these relationships can be used to derive queueing characteristics in steady state. We believe that this may both lead to new results and easier derivations of some known results.

The relationship in Section 7 has a different nature than the rest of this paper because it does not require any stationarity of the processes of interest. Namely, it suggests that such an asymptotic relationship may enable us to obtain queueing characteristics with some error bounds, not assuming any stationarity condition. This is completely different from the standard analysis in queueing theory. Thus, it would be interesting to see whether it can yield useful results for the performance evaluation of queueing models. We leave this for future studies.

Appendix

In the appendices below, we omit "a.s." because countably many events each of which occurs w.p. 1 simultaneously occur w.p. 1.

A Proof of Lemma 1

Since the proofs of (32) and (33) are similar, we only prove (32). Since π^e is a probability distribution, we can choose a sufficiently large a for each $\epsilon > 0$ such that

$$\sum_{\max(|x|,|y|)\geq a} \pi^e(\boldsymbol{x},\boldsymbol{y}) < \epsilon.$$

Let $\mathcal{S}_a = \{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{Z}_+^{2m}; max(|\boldsymbol{x}|,|\boldsymbol{y}|) < a\}$, then $\mathcal{S}_a$ is a finite set. Hence, summing both sides of (30) for $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{S}_a$ yields

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}_a} 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) = \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}_a} \pi^e(\boldsymbol{x},\boldsymbol{y}),$$

and therefore

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(x,y)\notin\mathcal{S}_a} 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) \\
&\qquad = 1 - \lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(x,y)\in\mathcal{S}_a} 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) \\
&\qquad = 1 - \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}_a} \pi^e(\boldsymbol{x},\boldsymbol{y}) = \sum_{\max(|\boldsymbol{x}|,|\boldsymbol{y}|)\geq a} \pi^e(\boldsymbol{x},\boldsymbol{y}) < \epsilon. \qquad (50)
\end{aligned}$$

Multiplying both sides of (30) by $g(\boldsymbol{x},\boldsymbol{y})$ and summing them for $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{S}_a$ yields

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(x,y)\in\mathcal{S}_a} g(\boldsymbol{x},\boldsymbol{y}) 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) = \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}_a} g(\boldsymbol{x},\boldsymbol{y})\pi^e(\boldsymbol{x},\boldsymbol{y}).$$

Let $\|g\| = \sup_{\boldsymbol{x},\boldsymbol{y}} g(\boldsymbol{x},\boldsymbol{y})$, which is finite by the assumption. Since (50) implies that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(x,y)\notin\mathcal{S}_a} g(\boldsymbol{x},\boldsymbol{y}) 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) \\
&\qquad \leq \|g\| \limsup_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \sum_{(\boldsymbol{x},\boldsymbol{y})\notin\mathcal{S}_a} 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) < \|g\|\epsilon, \\
&\sum_{(\boldsymbol{x},\boldsymbol{y})\notin\mathcal{S}_a} g(\boldsymbol{x},\boldsymbol{y})\pi^e(\boldsymbol{x},\boldsymbol{y}) < \|g\|\epsilon,
\end{aligned}$$

we have

$$\limsup_{n\to\infty} \Big| \frac{1}{n}\sum_{\ell=1}^{n} \sum_{x,y} g(\boldsymbol{x},\boldsymbol{y}) 1(\boldsymbol{X}(t_\ell^e-) = \boldsymbol{x}, \Delta\boldsymbol{N}^e(t_\ell^e) = \boldsymbol{y}) - \sum_{x,y} g(\boldsymbol{x},\boldsymbol{y})\pi^e(\boldsymbol{x},\boldsymbol{y}) \Big| < 2\|g\|\epsilon.$$

Letting $\epsilon \downarrow 0$, we arrive at (32).

B Proof of Lemma 2

In view of Lemma 1, it suffices to prove that, for $A \in 2^J \setminus \{\emptyset\}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} 1(\boldsymbol{X}(t_{A,\ell}^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_{A,\ell}^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^r(t_{A,\ell}^d) = \boldsymbol{z}) = \pi_A^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}). \tag{51}$$

It follows from v) that, for each $i \in J, \ell \geq 1, \boldsymbol{y} \in S_A, \boldsymbol{z} \in \mathbb{Z}_+^m$, there is a unique $k \geq 1$ such that $\ell \leq k$ and

$$\begin{aligned} 1(\boldsymbol{X}(t_{A,\ell}^d-) = \boldsymbol{x}, &\Delta \boldsymbol{N}^d(t_{A,\ell}^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^d(t_{A,\ell}^d) = \boldsymbol{z}) \\ &= 1(\boldsymbol{X}(t_k^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_k^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^d(t_k^d) = \boldsymbol{z}), \end{aligned}$$

and (29) and (40) imply

$$\lim_{t\to\infty} \frac{\tilde{N}_A^d(t)}{|\tilde{\boldsymbol{N}}^d|(t)} = \lim_{t\to\infty} \frac{\frac{1}{t}\tilde{N}_A^d(t)}{\frac{1}{t}|\tilde{\boldsymbol{N}}^d|(t)} = \frac{\lambda_A^d}{\lambda^d}.$$

Hence, for $\boldsymbol{y} \in S_A$,

$$\begin{aligned} \pi^d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 1(\boldsymbol{X}(t_k^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_k^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^r(t_k^d) = \boldsymbol{z}) \\ &= \lim_{t\to\infty} \frac{1}{|\tilde{\boldsymbol{N}}^d|(t)} \sum_{k=1}^{|\tilde{\boldsymbol{N}}^d|(t)} 1(\boldsymbol{X}(t_k^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_k^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^r(t_k^d) = \boldsymbol{z}) \\ &= \lim_{t\to\infty} \frac{\tilde{N}_A^d(t)}{|\tilde{\boldsymbol{N}}^d|(t)} \frac{1}{\tilde{N}_A^d(t)} \sum_{\ell=1}^{\tilde{\boldsymbol{N}}^d(t)} 1(\boldsymbol{X}(t_{A,\ell}^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_{A,\ell}^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^r(t_{A,\ell}^d) = \boldsymbol{z}) \\ &= \frac{\lambda_A^d}{\lambda^d} \lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} 1(\boldsymbol{X}(t_{A,\ell}^d-) = \boldsymbol{x}, \Delta \boldsymbol{N}^d(t_{A,\ell}^d) = \boldsymbol{y}, \Delta \boldsymbol{N}^r(t_{A,\ell}^d) = \boldsymbol{z}). \end{aligned}$$

This proves (51) by the definition π_A^e. The fact that π_A^d is a probability distribution is immediate from (43) with $h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \equiv 1$.

References

[1] F. Baccelli and P. Brémaud (2003). *Elements of queueing theory: Palm martingale calculus and stochastic recurrences*, vol. 26 of *Applications of Mathematics*. 2nd ed. Springer, Berlin.

[2] S.C. Borst and O.J. Boxma (1997). Polling models with and without switchover times. *Oper. Res.* **45**, 536-543.

[3] O.J. Boxma, O. Kella and K.M. Kosinski (2011). Queue lengths and workloads in polling systems. *Oper. Res. Letters* **39**, 401-405.

[4] M.A.A. Boon, R.D. van der Mei and E.M.M. Winands (2013). Waiting times in queueing networks with a single shared server. *Queueing Systems* **74**, 403-429.

[5] O.J. Boxma and T. Takine (2003). The $M/G/1$ FIFO queue with several customer classes. *Queueing Systems* **45**, 185-189.

[6] J.W. Cohen (1987). A two-queue, one-server model with priority for the longer queue. *Queueing Systems* **2**, 261-283.

[7] M. Eisenberg (1972). Queues with periodic service and changeover time. *Oper. Res.* **20**, 440-451.

[8] L. Flatto (1989). The longer queue model. *Prob. Eng. Inform. Sci.* **3**, 537-559.

[9] M. Miyazawa (1994) Rate conservation laws: a survey. *Queueing Systems* **15**, 1-58.

[10] M. Miyazawa (2010). *Palm calculus, reallocatable GSMP and insensitivity structure*, chap. 4 of Queueing networks: A fundamental approach. International Series in Operations Research and Management Science, Springer, 141–215.

[11] M. Sidi, H. Levy and S.W. Fuhrmann (1992). A queueing network with a single cyclically roving server. *Queueing Systems* **11**, 121-144.

[12] H. Takagi (1991). Queueing Analysis. A foundation of performance evaluation. Volume 1: Vacation and priority systems. *North-Holland Publ. Cy., Amsterdam*.

[13] Y.-S. Zheng and P. Zipkin (1990). A queueing model to analyze the value of centralized inventory information. *Oper. Res.* **38**, 296-307.