ASIP tandem queues with consumption

Yaron Yeger[a,*], Onno Boxma[a], Jacques Resing[a], Maria Vlasiou[a,b]

[a] *Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands*
[b] *University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands*

Abstract

The Asymmetric Inclusion Process (ASIP) tandem queue is a model of stations in series with a gate after each station. At a gate opening, all customers in that station instantaneously move to the next station unidirectionally. In our study, we enhance the ASIP model by introducing the capability for individual customers to independently move from one station to the next, and by allowing both individual customers and batches of customers from any station to exit the system. The model is inspired by the process by which macromolecules are transported within cells.

We present a comprehensive analysis of various aspects of the queue length in the ASIP tandem model. Specifically, we provide an exact analysis of queue length moments and correlations and, under certain circumstances, of the queue length distribution. Furthermore, we propose an approximation for the joint queue length distribution. This approximation is derived using three different approaches, one of which employs the concept of the replica mean-field limit. Among other results, our analysis offers insight into the extent to which nutrients can support the survival of a cell.

1. Introduction

This paper considers a broad class of queues in series, which contains the classical Jackson infinite server tandem queues as well as the Asymmetric Inclusion Process (ASIP) tandem queues as special cases. The ASIP (see [31, 32, 33, 34, 37, 38]) is a recent addition to the family of tandem stochastic systems (TSSs), which form an important class of stochastic networks that have found widespread use in various scientific fields.

A TSS comprises a linear stochastic network of relatively simple building blocks, with a stochastic input flow of particles progressing through a serial array of sites. Particles may represent customers, messages, products, calls, jobs, molecules, nutrients, etc. TSSs are governed by a set of rules characterizing the system's law of motion and often exhibit complex stochastic dynamics. While the isolated behavior of each building block may be well understood, predicting the behavior of the aggregate can be incredibly challenging. Despite the inherent complexity of TSS analysis, it remains a vital area of study for a host of scientific communities.

Tandem Jackson Networks (TJNs) and the Asymmetric Exclusion Process (ASEP) are two notable examples of TSSs. TJNs are a sequential array of n service stations or sites. External particles arrive at station 1 according to a Poisson process. Each station has an unlimited buffer size, and particles individually move from station i to station $i+1$, leaving the system from station n. Under certain conditions on the service disciplines at the stations and the service time distributions at each station, the number of customers at each station are independent, resulting in a product form of the joint queue length distribution. After the pioneering work of R.R.P. Jackson [24] and J.R. Jackson [22, 23], product-form networks were studied extensively, with key contributions in [3, 27] and major applications in computer-communications; see, e.g., [7, 14].

*Corresponding author
Email addresses: yaronyeg@mail.tau.ac.il (Yaron Yeger), o.j.boxma@tue.nl (Onno Boxma), j.a.c.resing@tue.nl (Jacques Resing), m.vlasiou@utwente.nl (Maria Vlasiou)

Contrary to TJN, ASEP is a TSS where each site can hold at most a single particle, causing blocking on the (forward) movement of particles. ASEP is a representation of a one-way driven lattice gas comprising particles that are subject to exclusion interactions. A gas–liquid system can be modelled by means of hard-core particles, which exclude one another within a non-zero range. A further simplification consists of the restriction of the particle coordinates to the vertices of a regular lattice. Moreover, the ratio of the hard-core radius to the lattice constant can be chosen such that the exclusion is restricted to nearest-neighbor sites, represented by the ASEP model. This results in congestion throughout the system, making its dynamics extremely complex. ASEP has been extensively studied, in particular, in the physics literature [17, 20, 35].

ASIP is a TSS that 'closes the divide' between TJNs and ASEP. ASIP is referred to as the "bosonic" counterpart of ASEP since the exclusion principle that sets particles apart in the ASEP is replaced by the inclusion principle that causes them to form inseparable clusters in ASIP, while particles in both ASEP and ASIP move unidirectionally along a one-dimensional lattice due to random events. The ASIP's irreversible tendency for particles to stick together makes it suitable for modeling physical systems whose behavior goes against ASEP's exclusion principle. ASIP serves, e.g., as a lattice-gas model for unidirectional transport with irreversible aggregation. In this model's dynamics, particles move in a unidirectional manner and each site can accommodate an arbitrary number of particles simultaneously. The inclusion principle allows these particles to form clusters that move together to the next site. ASIP is characterized by an unbounded buffer capacity and unlimited batch service. This means that at completion of service at a site all particles present at that site move as a cluster to the next site. This model can be viewed as a tandem array of growth-collapse processes. Additionally, it finds application in various fields such as road-traffic analysis, where vehicles move forward to the next traffic light when it turns green, and marine traffic analysis along a canal with locks, like the Panama Canal; see [32] for further information and references therein.

Goal and motivation. The goal of the present paper is to study the performance of a tandem ASIP with the following additional feature: next to the cluster movements when a gate opens, we also allow individual particles to move either to the next queue or out of the system. Our motivation for this is twofold. Firstly, we aim to develop and analyze a very general tandem queue that contains both the original ASIP as well as $M/M/\infty$ queues in series as special cases. The model thus holds promising potential for utilization within the field of computer-communications, as well as in domains such as road traffic and physical chemistry. Secondly, we would like to enhance the ASIP model in such a way that it can represent the so-called *Vesicular Transport Hypothesis*. This theory explains how macromolecules, such as proteins, are transported within cells [6]. During the process of vesicular transport, some of the proteins that are transported within cells can be *consumed* and *degraded*, motivating our extension to including consumption as a new feature to the classical ASIP model.

The Vesicular Transport Hypothesis theory proposes that transport vesicles, small membrane-bound sacs, move cargo between different compartments of a eukaryotic cell, including the endoplasmic reticulum, Golgi apparatus, lysosomes, and plasma membrane. According to the Vesicular Transport Hypothesis, cargo molecules are packaged into transport vesicles at their site of synthesis and then transported to the Golgi apparatus, where the cargo is sorted and modified. New vesicles then bud off and transport the cargo to its final destination. This hypothesis has been supported by numerous experiments, including the visualization of transport vesicles under electron microscopy and the identification of the proteins that regulate vesicle formation, fusion, and movement [11, 21, 29, 36]. Overall, vesicular transport is a dynamic process that involves the movement of proteins between different compartments of the cell, and some of these proteins may be consumed or secreted during the process.

While in mathematical biology, microbial cell factories is an established approach in bioengineering that considers a cell as a production facility, the current conventional techniques (which notably led to CRISPR and earlier technologies such as Zinf-finger nucleases) are very time consuming, are hard to analyze, and typically destroy the cells that is experimented upon. An emerging area in mathematical biology is to use the experience built in OR in actual production

facilities, by proposing suitable queueing models that support all existing biological experiments for a specific mechanism. In particular, for the Vesicular Transport Hypothesis theory, observations do show the general behavior of the arrival and transportation of nutrients to and from cells. However, due to the complexity of this process a stochastic mathematical modeling of it was not analyzed prior, yielding explicit results.

The ASIP model can serve as a mathematical representation of the process of vesicular transport and transport between cells [25]. However, a new property should be introduced to the model, namely, the consumption or degradation of particles in sites during their stay. Our proposed framework expands the movement of particles between sites in the ASIP model, beyond the conventional cluster movement, to include individual movement between sites and consumption within sites. By introducing these additional options, our framework offers a more comprehensive analysis of the ASIP model, enabling the exploration of a wider range of system behaviors in the family of TSSs with unlimited storage capacity per site, bridging the gap between the two extreme cases represented by the TJN and classical ASIP models. Including a form of consumption in the ASIP system is a crucial step towards representing the Vesicular Transport Hypothesis. We obtain exact moment and correlation results for an n-station ASIP with two different forms of consumption. One of the relevant questions we answer in the paper is how far down-stream, can nutrients still sustain a cell.

Main contributions. The main contributions of the paper are: (i) We present a very general framework for TSSs with cluster movements as well as individual movements. (ii) We show how (joint) moments of any order can be obtained in a recursive way. (iii) We present an exact analysis of the queue length distribution for $n = 1$ station, and we propose an approximation approach for queue length distributions for general n.

We use three different reasonings to arrive at one and the same approximation; one of them is the recently developed Replica Mean-Field limit approach [2]. The approximation gives exact mean queue lengths at all queues. We furthermore show via asymptotics and numerical plots that, for a wide parameter range, second moments and correlations of queue lengths are quite accurately approximated. However, it should be realized that the model contains many parameters which may be totally different when switching, say, from a road traffic application to a biological cell application. Hence, dependent on the specific application, one might focus on different approximations or asymptotic regimes.

Organization of the paper. In Section 2, we present and analyze the extended ASIP model, where consumption is being defined by two different parallel processes: (i) binomial consumption; and (ii) individual consumption. In Sections 3 and 4, we analyze each of these separate cases. Section 5 contains conclusions and some suggestions for further research.

2. The general model

In this section, we consider the general ASIP tandem model with consumption. We provide the mathematical formulation of the model in Section 2.1. The balance equations for the steady-state distribution of the numbers of particles/customers at all stations are presented in Section 2.2, leading to a functional equation for the probability generating function (PGF) of that distribution. Last, in Section 2.3, we show how that functional equation can be used to obtain moments and cross-correlations for the numbers of particles at each site.

2.1. Model description

The model under consideration is a system consisting of n stations (or queues) in series, with a Poisson(λ) arrival process at station 1 and with a gate after each station. Each gate opens after i.i.d. exponentially distributed intervals, at rate μ_i for gate i, $i = 1, 2, \ldots, n$. When gate i opens, all particles in station i instantaneously move to station $i + 1$, $i = 1, 2, \ldots, n - 1$ and the gate immediately closes again; when gate n opens, the particles in station n leave the system. We represent the consumption of particles (or, depending on the application: leakage, abandonment,

or service of an individual particle that subsequently leaves the system) in two different ways. Firstly, at i.i.d. $\exp(\tau_i)$ distributed intervals, each individual particle in station i leaves the system with probability $1-a_i$ and stays in station i with probability a_i, $i=1,\ldots,n$. Hence, if station i contains j_i particles at such an event, a binomially distributed number, with parameters j_i and $1-a_i$, leaves station i. Secondly, each particle in station i is consumed after an exponentially distributed amount of time with rate $(1-p_i)\nu_i$, $i=1,2,\ldots,n$. In addition, we assume that each particle in station i moves to station $i+1$ after an exponentially distributed time with rate $p_i\nu_i$, $i=1,2,\ldots,n$, where station $n+1$ is interpreted as 'out of the system'. Of course, we could have introduced some rates θ_i and ζ_i instead of $(1-p_i)\nu_i$ and $p_i\nu_i$, but the two different move events of an individual particle of station i will give rise to very similar expressions for the PGF, which we shall combine. Finally, all interarrival times at station 1 and all consumption intervals and gate opening intervals at all stations are assumed to be independent.

Particular parameter choices lead to well-studied models. In particular, taking all ν_i and τ_i equal to zero results in the tandem ASIP system studied by Reuveni et al. [32]; and taking all $\mu_i, 1-p_i$ and τ_i equal to zero results in a tandem system of $M/M/\infty$ queues (cf. [23, 3]; and see [8] for the transient joint distribution of queue lengths and residual service times at all stations, for the case of generally distributed service times). We observe that allowing additional independent Poisson arrival processes at stations $2,\ldots,n$ does not complicate the analysis significantly.

We consider the distribution of particles over the stations in steady state; a steady-state distribution obviously exists in view of the unlimited capacity of the gates to transport particles.

2.2. Balance equations

Let $p(j_1,j_2,\ldots,j_n)$ denote the steady state probability of having $j_1,j_2,\ldots,j_n$ particles in stations $1,2,\ldots,n$. It is readily verified that the balance equations for the case of n stations in series are given by (with $I(\cdot)$ an indicator function):

$$\begin{aligned}
[\lambda+\sum_{i=1}^{n} j_i\nu_i+\sum_{i=1}^{n}\tau_i+\sum_{i=1}^{n}\mu_i]p(j_1,\ldots,j_n) &= \lambda p(j_1-1,j_2,\ldots,j_n)I(j_1>0)\\
&+\sum_{i=1}^{n}(j_i+1)\nu_i[(1-p_i)p(j_1,\ldots,j_{i-1},j_i+1,j_{i+1},\ldots,j_n)\\
&+p_i p(j_1,\ldots,j_{i-1},j_i+1,j_{i+1}-1,\ldots,j_n)I(j_{i+1}>0)]\\
&+\sum_{i=1}^{n}\tau_i\sum_{k=0}^{\infty}p(j_1,\ldots,j_i+k,j_{i+1},\ldots,j_n)\binom{k+j_i}{k}a_i^{j_i}(1-a_i)^k\\
&+\sum_{i=1}^{n-1}\mu_i\sum_{k=0}^{j_{i+1}}p(j_1,\ldots,j_{i-1},k,j_{i+1}-k,j_{i+2},\ldots,j_n)I(j_i=0)\\
&+\mu_n\sum_{k=0}^{\infty}p(j_1,\ldots,j_{n-1},k)I(j_n=0). \qquad (1)
\end{aligned}$$

Let $X_1,\ldots,X_n$ denote the steady-state numbers of particles in stations $1,\ldots,n$. Denote their PGF as

$$P(z_1,\ldots,z_n) := \mathbb{E}[z_1^{X_1}\ldots z_n^{X_n}].$$

It easily follows from (1) that

$$
\begin{aligned}
&[\lambda + \sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} \mu_i] P(z_1, \ldots, z_n) + \sum_{i=1}^{n} \nu_i z_i \frac{\partial}{\partial z_i} P(z_1, \ldots, z_n) \\
&= \lambda z_1 P(z_1, \ldots, z_n) + \sum_{i=1}^{n-1} [(1-p_i)\nu_i + p_i \nu_i z_{i+1}] \frac{\partial}{\partial z_i} P(z_1, \ldots, z_n) + \nu_n \frac{\partial}{\partial z_n} P(z_1, \ldots, z_n) \\
&+ \sum_{i=1}^{n} \tau_i P(z_1, \ldots, z_{i-1}, a_i z_i + 1 - a_i, z_{i+1}, \ldots, z_n) \\
&+ \sum_{i=1}^{n-1} \mu_i P(z_1, \ldots, z_{i-1}, z_{i+1}, z_{i+1}, z_{i+2}, \ldots, z_n) + \mu_n P(z_1, \ldots, z_{n-1}, 1), \qquad (2)
\end{aligned}
$$

and hence

$$
\begin{aligned}
&[\lambda(1-z_1) + \sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} \mu_i] P(z_1, \ldots, z_n) \\
&= \sum_{i=1}^{n-1} \nu_i [(1-p_i)(1-z_i) + p_i(z_{i+1} - z_i)] \frac{\partial}{\partial z_i} P(z_1, \ldots, z_n) + \nu_n (1-z_n) \frac{\partial}{\partial z_n} P(z_1, \ldots, z_n) \\
&+ \sum_{i=1}^{n} \tau_i P(z_1, \ldots, z_{i-1}, a_i z_i + 1 - a_i, z_{i+1}, \ldots, z_n) \\
&+ \sum_{i=1}^{n-1} \mu_i P(z_1, \ldots, z_{i-1}, z_{i+1}, z_{i+1}, z_{i+2}, \ldots, z_n) + \mu_n P(z_1, \ldots, z_{n-1}, 1). \qquad (3)
\end{aligned}
$$

2.3. Moments

Define

$$
m_{(k_1,k_2,\ldots,k_n)} = \frac{d^{k_1}}{dz_1^{k_1}} \cdots \frac{d^{k_n}}{dz_n^{k_n}} P(z_1, z_2, \ldots, z_n)_{|z_1 = z_2 = \cdots = z_n = 1}. \qquad (4)
$$

Then we have from (3), by taking derivatives with respect to all the variables and subsequently taking $z_1 = \cdots = z_n = 1$, that

$$
\begin{aligned}
\left(\sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} \mu_i \right) m_{(k_1,k_2,\ldots,k_n)} &= k_1 \lambda m_{(k_1-1,k_2,\ldots,k_n)} I(k_1 > 0) - \left(\sum_{i=1}^{n} k_i \nu_i \right) m_{(k_1,k_2,\ldots,k_n)} \\
&+ \sum_{i=1}^{n-1} p_i \nu_i k_{i+1} m_{(k_1,\ldots,k_{i-1},k_i+1,k_{i+1}-1,k_{i+2},\ldots,k_n)} I(k_{i+1} > 0) + \left(\sum_{i=1}^{n} \tau_i a_i^{k_i} \right) m_{(k_1,k_2,\ldots,k_n)} \\
&+ \sum_{i=1}^{n-1} \mu_i \sum_{j=0}^{k_{i+1}} \binom{k_{i+1}}{j} m_{(k_1,\ldots,k_{i-1},j,k_{i+1}-j,k_{i+2},\ldots,k_n)} I(k_i = 0) + \mu_n m_{(k_1,k_2,\ldots,k_n)} I(k_n = 0). \qquad (5)
\end{aligned}
$$

Equation (5) can be used to recursively find (joint) factorial moments of the random variables $X_1, X_2, \ldots X_n$. In the sequel, we illustrate this for a number of cases. Let us start with all the factorial moments of the random variable X_1. Define $m_1^{(k)} = m_{(k,0,\ldots,0)} = \mathbb{E}\left(\prod_{i=0}^{k-1}(X_1 - i)\right)$, with by definition $m_1^{(0)} = 1$. From (5), we have that

$$
m_1^{(k)} = \frac{k\lambda m_1^{(k-1)}}{\mu_1 + k\nu_1 + \tau_1(1 - a_1^k)}, \quad k = 1, 2, \ldots, \qquad (6)
$$

and hence

$$
m_1^{(k)} = \frac{k! \lambda^k}{\prod_{j=1}^{k} \left(\mu_1 + j\nu_1 + \tau_1(1 - a_1^j) \right)}. \qquad (7)
$$

Next we obtain the first moment of all the random variables $X_1, X_2, \ldots, X_n$. We already have from (7) that

$$\mathbb{E}X_1 = \frac{\lambda}{\mu_1 + \nu_1 + \tau_1(1-a_1)}. \tag{8}$$

Taking $k_j = 1$ and $k_i = 0$ for $i \neq j$ in (5) yields

$$(\mu_j + \nu_j + \tau_j(1-a_j))\mathbb{E}X_j = (\mu_{j-1} + \nu_{j-1}p_{j-1})\mathbb{E}X_{j-1}, \quad j = 2, \ldots, n, \tag{9}$$

and hence

$$\mathbb{E}X_i = \frac{\lambda}{\mu_1 + \nu_1 + \tau_1(1-a_1)} \prod_{j=2}^{i} \frac{\mu_{j-1} + \nu_{j-1}p_{j-1}}{\mu_j + \nu_j + \tau_j(1-a_j)}, \quad i = 2, \ldots, n. \tag{10}$$

Remark 1. *Observe that (8) (rewritten as $(\mu_1 + \nu_1 + \tau_1(1-a_1))\mathbb{E}X_1 = \lambda$) and (9) are flow balance equations: they equate the steady-state flow out of a station and the steady-state flow into that same station.*

Remark 2. *If $\mu_j = \mu$, $\nu_j = \nu$, $p_j = p$, $\tau_j = \tau$ and $a_j = a$ for all $j = 1, \ldots, n$, then $\mathbb{E}X_i = \frac{\lambda}{\mu+\nu+\tau(1-a)}(\frac{\mu+\nu p}{\mu+\nu+\tau(1-a)})^{i-1}$. If the mean number of particles should stay above a critical value C for a cell to be in good shape, then apparently i should be smaller than $\ln(\lambda/(C[\mu + \nu p]))/\ln(\frac{\mu+\nu+\tau(1-a)}{\mu+\nu p})$.*

We next turn to the (joint) factorial moments $\mathbb{E}X_i(X_i - 1)$ and $\mathbb{E}X_iX_j$. We already have from (6) that

$$\mathbb{E}X_1(X_1 - 1) = \frac{2\lambda\mathbb{E}X_1}{\mu_1 + 2\nu_1 + \tau_1(1-a_1^2)}, \tag{11}$$

and hence

$$\mathrm{Var}(X_1) = \mathbb{E}X_1[\frac{\lambda\mu_1 + \lambda\tau_1(1-a_1)^2}{[\mu_1 + 2\nu_1 + \tau_1(1-a_1^2)][\mu_1 + \nu_1 + \tau_1(1-a_1)]} + 1]. \tag{12}$$

Furthermore, (5) with $n = 2$ and $k_1 = k_2 = 1$ yields:

$$\begin{aligned}
\mathbb{E}X_1X_2 &= \frac{\lambda\mathbb{E}X_2 + \nu_1p_1\mathbb{E}X_1(X_1-1)}{\mu_1 + \nu_1 + \tau_1(1-a_1) + \mu_2 + \nu_2 + \tau_2(1-a_2)} \\
&= \frac{\lambda^2}{[\mu_1 + \nu_1 + \tau_1(1-a_1) + \mu_2 + \nu_2 + \tau_2(1-a_2)][\mu_1 + \nu_1 + \tau_1(1-a_1)]} \\
&\quad \times \left[\frac{\mu_1 + \nu_1p_1}{\mu_2 + \nu_2 + \tau_2(1-a_2)} + \frac{2\nu_1p_1}{\mu_1 + 2\nu_1 + \tau_1(1-a_1^2)}\right].
\end{aligned} \tag{13}$$

We finally obtain the following expression for the covariance of X_1 and X_2:

$$\begin{aligned}
\mathrm{cov}(X_1, X_2) &= \frac{1}{\mu_1 + \nu_1 + \tau_1(1-a_1) + \mu_2 + \nu_2 + \tau_2(1-a_2)} \\
&\quad \times \left[\frac{-\lambda^2(\mu_1 + \nu_1p_1)}{(\mu_1 + \nu_1 + \tau_1(1-a_1))^2} + \frac{2\lambda^2\nu_1p_1}{(\mu_1 + \nu_1 + \tau_1(1-a_1))(\mu_1 + 2\nu_1 + \tau_1(1-a_1^2))}\right].
\end{aligned} \tag{14}$$

Taking $n = 2$, $k_1 = 0$ and $k_2 = 2$ in (5) yields

$$[\mu_2 + 2\nu_2 + \tau_2(1-a_2^2)]\mathbb{E}X_2(X_2-1) = \mu_1\mathbb{E}X_1(X_1-1) + 2(\mu_1 + \nu_1p_1)\mathbb{E}X_1X_2, \tag{15}$$

where $\mathbb{E}X_1(X_1 - 1)$ is given in (11) and $\mathbb{E}X_1X_2$ in (13). $\mathrm{Var}(X_2)$ subsequently follows by adding $\mathbb{E}X_2 - (\mathbb{E}X_2)^2$ to the expression for $\mathbb{E}X_2(X_2 - 1)$ obtained in (15). Taking $k_i = 2$, for arbitrary $i \geq 2$, and $k_j = 0$ for all $j \neq i$ in (5), we obtain:

$$\begin{aligned}
&[\mu_i + 2\nu_i + \tau_i(1-a_i^2)]\mathbb{E}X_i(X_i-1) \\
&\qquad = 2(\mu_{i-1} + \nu_{i-1}p_{i-1})\mathbb{E}X_{i-1}X_i + \mu_{i-1}\mathbb{E}X_{i-1}(X_{i-1}-1).
\end{aligned} \tag{16}$$

Taking $k_i = k_j = 1$ for arbitrary i and j with $j > i$ in (5) and furthermore $k_\ell = 0$ for $\ell \neq i$ and $\ell \neq j$, we obtain:

$$\begin{aligned}[\mu_i + \nu_i + \tau_i(1-a_i) + \mu_j + \nu_j + \tau_j(1-a_j)]\mathbb{E}X_iX_j \\ = (\mu_{i-1} + \nu_{i-1}p_{i-1})\mathbb{E}X_{i-1}X_j\ I(i>1) + (\mu_{j-1} + \nu_{j-1}p_{j-1})\mathbb{E}X_iX_{j-1}\ I(j>i+1) \\ + \nu_i p_i \mathbb{E}X_i(X_i-1)\ I(j=i+1) + \lambda\mathbb{E}X_j\ I(i=1). \quad (17)\end{aligned}$$

These relations clearly show how all $\mathrm{Var}(X_i)$ and $\mathrm{cov}(X_i, X_j)$ can be derived recursively, starting from $\mathrm{Var}(X_1)$ and $\mathrm{cov}(X_1, X_2)$.

Remark 3. *Observe that Fiems et al. [19] obtain moments for a network that contains the present tandem system as a special case, except that binomial transitions do not fit in their model; they present equations from which transient moments can be obtained, but they do not give explicit expressions for stationary moments.*

While it would be very interesting to explicitly determine $P(z_1, \ldots, z_n)$, that seems to be prohibitively hard for the general model under consideration in this section. For $n = 1$, with $P(z)$ the PGF of the steady-state number of particles in station 1, Equation (3) becomes

$$[\lambda(1-z) + \tau_1 + \mu_1]P(z) = \nu_1(1-z)\frac{\mathrm{d}}{\mathrm{d}z}P(z) + \tau_1 P(a_1 z + 1 - a_1) + \mu_1. \quad (18)$$

This equation corresponds to the equation for the PGF of the steady-state number of customers in the $M/M/\infty$ model with both binomial and total catastrophes. This model is a special case of the linear birth/immigration-death process with binomial catastrophes considered in Kapodistria et al. [26]. There the authors give an expression for the factorial moments of the steady-state number of customers and an expression for the steady-state probabilities in terms of these factorial moments. See in particular Equations (7.3), (7.4) and (7.5) in Section 7.1 in [26] where they consider two or more binomial catastrophes (one of them representing the total catastrophes in our case).
Observe that Equation (18), for the 1-station system only, is an inhomogeneous differential-delay equation. In the next two sections, we turn our attention to two special cases: (i) the case in which $\nu_i = 0\ \forall i$ (which we call 'binomial consumption') and (ii) the case $\tau_i = 0\ \forall i$ (which we call 'individual consumption'). For these two cases, we first determine an explicit expression for the PGF $P(z)$ for the 1-station system. The 2-station model with either $\tau_1 = \tau_2 = 0$ or $\nu_1 = \nu_2 = 0$ gives rise to a boundary value problem (cf. [16, 18]) that seems too complicated; however, we propose an approximation for both models that is shown to be very accurate for a large parameter range and that allows an extension to the n-station model.

3. Binomial consumption

In this section, we consider the general model, with one exception: all ν_i are assumed to be zero. In Section 3.1, we determine the PGF of the number of particles in the first station. Section 3.2 discusses the model with two stations in series, suggesting three different approximation ideas – which turn out to lead to the same approximation – for the PGF of the steady-state joint distribution of particles over the two stations. In Section 3.3, we show how this approximation can be extended to the n-station case.

3.1. The case $n = 1$

In this section, we determine the *distribution* of the number of particles in station 1 in steady state. Taking $\nu_1 = 0$ in (18) shows that the PGF $P(z) \equiv P(z, 1, \ldots, 1)$ satisfies the following recurrence relation:

$$[\lambda(1-z) + \tau_1 + \mu_1]P(z) = \tau_1 P(a_1 z + 1 - a_1) + \mu_1. \quad (19)$$

Introducing $q_1 := \frac{\tau_1}{\tau_1+\mu_1}$ and $g(z) := \frac{\tau_1+\mu_1}{\tau_1+\mu_1+\lambda(1-z)}$, we rewrite (19) into

$$P(z) = q_1 g(z) P(a_1 z + 1 - a_1) + (1 - q_1) g(z). \tag{20}$$

We solve this equation by iteration, with as first step, defining $f(z) := a_1 z + 1 - a_1$:

$$P(z) = (1 - q_1) g(z) + q_1 g(z) \left[(1 - q_1) g(f(z)) + q_1 g(f(z)) P(f(f(z))) \right]. \tag{21}$$

Introducing $f^{(0)}(z) := z$ and $f^{(j)}(z) = f(f^{(j-1)}(z))$ for $j = 1, 2, \ldots$, it is quickly seen that $f^{(j)}(z) = a_1^j z + 1 - a_1^j$, and that continued iteration of (21) results in

$$P(z) = \lim_{M\to\infty} [\sum_{j=0}^{M} (1-q_1) q_1^j \prod_{i=0}^{j} g(a_1^i z+1-a_1^i) + P(a_1^{M+1} z+1-a_1^{M+1}) q_1^{M+1} \prod_{i=0}^{M} g(a_1^i z+1-a_1^i). \tag{22}$$

The last term clearly converges to zero as all $g(\cdot)$ terms are bounded by one and $P(a_1^{j+1} z+1-a_1^{j+1})$ tends to $P(1) = 1$. For the case of $n = 1$ and no individual consumption, we have thus proven that the PGF of the steady-state number of particles is given by

$$P(z) = \sum_{j=0}^{\infty} \frac{\mu_1}{\tau_1 + \mu_1} (\frac{\tau_1}{\tau_1 + \mu_1})^j \prod_{i=0}^{j} \frac{\tau_1 + \mu_1}{\tau_1 + \mu_1 + \lambda a_1^i (1 - z)}. \tag{23}$$

This result can be easily interpreted, realizing that $g(z)$ is the PGF of the number of Poisson(λ) arrivals during an interval that is exponentially distributed with rate $\tau_1 + \mu_1$. With probability $\frac{\mu_1}{\tau_1+\mu_1}(\frac{\tau_1}{\tau_1+\mu_1})^j$, there are exactly j consumption epochs between two successive gate openings. And in that event, arrivals in all the $j + 1$ intervals between two successive gate openings are binomially thinned with probabilities a_1^j (for those arriving before the first consumption), $a_1^{j-1}, \ldots, a_1^0$, resulting in those PGF's $\frac{\tau_1+\mu_1}{\tau_1+\mu_1+\lambda a_1^i(1-z)}$.

Either from the expression for $P(z)$ or from (19), we readily obtain the mean and variance of X_1 (cf. also (8), (11) and (12)):

$$\mathbb{E}X_1 = \frac{\lambda}{\mu_1 + \tau_1(1 - a_1)}, \tag{24}$$

$$\mathbb{E}X_1(X_1 - 1) = \frac{2\lambda^2}{(\mu_1 + \tau_1(1 - a_1))(\mu_1 + \tau_1(1 - a_1^2))}, \tag{25}$$

and hence

$$\mathrm{Var}(X_1) = \frac{\lambda}{\mu_1 + \tau_1(1 - a_1)} [\frac{\lambda\mu_1 + \lambda\tau_1(1 - a_1)^2}{(\mu_1 + \tau_1(1 - a_1))(\mu_1 + \tau_1(1 - a_1^2))} + 1]. \tag{26}$$

3.2. The case $n = 2$

In this section, we discuss the general model with two stations, but without the individual consumption: $\nu_1 = \nu_2 = 0$. It follows from (3) that the two-dimensional PGF of the steady-state joint distribution of numbers of particles (X_1, X_2) in both stations now is given by

$$\begin{aligned}[\lambda(1 - z_1) + \tau_1 + \tau_2 + \mu_1 + \mu_2] P(z_1, z_2) \\ = \tau_1 P(a_1 z_1 + 1 - a_1, z_2) + \tau_2 P(z_1, a_2 z_2 + 1 - a_2) + \mu_1 P(z_2, z_2) + \mu_2 P(z_1, 1).\end{aligned} \tag{27}$$

The $P(z_2, z_2)$ term makes this recursion very difficult to solve; this even holds for the one-dimensional equation that arises by taking $z_1 = 1$:

$$[\tau_2 + \mu_1 + \mu_2] P(1, z_2) = \tau_2 P(1, a_2 z_2 + 1 - a_2) + \mu_1 P(z_2, z_2) + \mu_2. \tag{28}$$

Below we suggest three approximations for the marginal PGF of X_2; they will turn out to amount to the same.

Approximation 1. The arrival process at station 2 is a Poisson(μ_1) process of batches. By PASTA (Poisson Arrivals See Time Averages), the PGF of the number of particles in station 1 just before a gate opening is the same as the PGF of the steady-state number of particles $P(z)$. Hence, the batch size PGF equals $P(z)$. However, the size of a batch entering station 2 is correlated with the length of the preceding gate opening interval. The approximation that we propose is to ignore that dependence.

Let $\pi(j)$ denote the steady-state probability of having j particles in the resulting infinite server queue, and denote the (exact) probability of a batch of size m coming from station 1 by B_m. It is easily seen that the balance equations in the resulting infinite server queue with batch arrivals are given by

$$[\mu_1+\tau_2+\mu_2]\pi(j)=\mu_1\sum_{i=0}^{j}\pi(i)B_{j-i}+\tau_2\sum_{k=0}^{\infty}\pi(j+k)\binom{j+k}{k}a_2^j(1-a_2)^k+\mu_2 I(j=0). \quad (29)$$

Taking generating functions, with $P^{app}(1,z)$ the PGF of $\pi(j)$, we obtain the following equation:

$$[\mu_1(1-P(z))+\tau_2+\mu_2]P^{app}(1,z)=\tau_2 P^{app}(1,a_2z+1-a_2)+\mu_2. \quad (30)$$

This equation has exactly the same structure as (19), but with the term $\lambda(1-z)$ replaced by $\mu_1(1-P(z))$. Introducing $f(z):=a_2z+1-a_2$, so that the j-th iterate $f^{(j)}(z)=a_2^jz+1-a_2^j$ for $j=0,1,\ldots$, and following exactly the same iteration procedure as in the previous section, we obtain the following expression for the approximation of $P(1,z)$:

$$P^{app}(1,z)=\sum_{j=0}^{\infty}\frac{\mu_2}{\tau_2+\mu_2}\left(\frac{\tau_2}{\tau_2+\mu_2}\right)^j\prod_{i=0}^{j}\frac{\tau_2+\mu_2}{\tau_2+\mu_2+\mu_1(1-P(a_1^iz+1-a_1^i))}. \quad (31)$$

The interpretation is the same as the one of Equation (23), with now $g(z)=\frac{\tau_1+\mu_1}{\tau_1+\mu_1+\lambda(1-z)}$ being replaced by $\frac{\tau_2+\mu_2}{\tau_2+\mu_2+\mu_1(1-P(z))}$, the PGF of the number of batch arrivals during an interval that is exponentially distributed with rate $\tau_2+\mu_2$.

Approximation 2. An alternative approach is to assume that X_2 is independent of X_1 – an independence that would hold when $\mu_1=0$ and also when $\mu_1\to\infty$. Under the independence assumption we have $P(z_1,z_2)=P(z_1,1)P(1,z_2)$, and hence (28) again reduces to (30), with $P(z_2)=P(z_2,1)$ given in (23).

Approximation 3. We now outline yet another approach for obtaining approximations for the ASIP model with consumption, viz., the Replica Mean-Field (RMF) limit approach. It emerges that the ordinary mean-field limit approach does not lead to a useful approximation, while RMF corroborates the other two approximations we propose. RMF was first suggested for different models in [2]; see also [1].

To address the issue of correlations and dependencies between different sites in the system, we replicate the ASIP tandem queue R times to create parallel tandem queues with identical intrinsic properties. Specifically, we consider R replicas of a 2-site ASIP tandem queue with consumptions, indexed by r ($1\le r\le R$). The arrival process to each replica is a Poisson(λ) arrival process at the first site. Gate j in each replica opens after independent and identically, exponentially, distributed intervals at rate μ_j. When the gate of the first site in replica r opens, all particles in that site instantaneously move to the second site in replica s ($1\le s\le R$), with equal probabilities $p_{rs}=\frac{1}{R}$. When the gate of the second site in replica r opens, all particles in that site instantaneously exit the replica, for all $r=1,2,\ldots,R$. Throughout the evolution of the replicas, particles are being consumed in each replica, as described in the introduction and in previous sections.

Denote by X_j^r the number of particles in site j in the r^{th} replica. Focusing on transitions that influence the states of the r^{th} replica, the Markovian dynamics in the RMF system leads to

$$(\lambda+\tau_1+\tau_2+\mu_1(1+\frac{R-1}{R})+\mu_2)\mathbb{E}[z_1^{X_1^r}z_2^{X_2^r}]$$
$$=\lambda z_1\mathbb{E}[z_1^{X_1^r}z_2^{X_2^r}]+\frac{\mu_1}{R}\mathbb{E}[z_2^{X_1^r+X_2^r}]+\frac{\mu_1(R-1)}{R}\mathbb{E}[z_2^{X_2^r}]+\sum_{s=1,s\neq r}^{R}\frac{\mu_1}{R}\mathbb{E}[z_1^{X_1^r}z_2^{X_1^s+X_2^r}]+\mu_2\mathbb{E}[z_1^{X_1^r}]$$
$$+\tau_1\mathbb{E}[(a_1z_1+1-a_1)^{X_1^r}z_2^{X_2^r}]+\tau_2\mathbb{E}[z_1^{X_1^r}(a_2z_2+1-a_2)^{X_2^r}]. \quad (32)$$

Note that we get a rate of $\mu_1(1+\frac{R-1}{R})$ on the left-hand side because gate openings in other replicas influence the state of the r^{th} replica with probability $\frac{1}{R}$. The second and third terms on the right-hand side represent transitions due to gate opening at site 1 in replica r. The second term represents a transfer of particles within the replica itself, whereas the third term represents a transfer of particles into other replicas. The fourth term on the right-hand side relates to gate openings in other replicas with transfer of particles into replica r.

Now, introduce the notation $P_r(z_1,z_2)=\mathbb{E}[z_1^{X_1^r}z_2^{X_2^r}]$ and $P_s(z_1,z_2)=\mathbb{E}[z_1^{X_1^s}z_2^{X_2^s}]$. Taking the limit $R\to\infty$, we assume that the dynamics of replicas become asymptotically independent (this is usually referred to as the 'Poisson hypothesis'). This yields $\mathbb{E}[z_1^{X_1^r}z_2^{X_1^s+X_2^r}]=P_r(z_1,z_2)P_s(z_2,1)$ and using the fact that $P_s(z_2,1)=P_r(z_2,1)$, we obtain from Equation (32) the following approximation:

$$((1-z_1)\lambda+\tau_1+\tau_2+\mu_1(2-P_r(z_2,1))+\mu_2)P_r(z_1,z_2)$$
$$=\mu_1P_r(1,z_2)+\mu_2P_r(z_1,1)+\tau_1P_r(a_1z_1+1-a_1,z_2)+\tau_2P_r(z_1,a_2z_2+1-a_2). \quad (33)$$

Taking $z_1=1$ reduces this last formula to Equation (30), which makes sense: replicating the tandem queues infinitely many times and having equal routing probabilities from all replicas of station 1 to all replicas of station 2 basically amounts to having Poisson arrivals of batches, where the batch sizes become independent of the gate opening interval at our station 1; it also amounts to independence of the numbers of particles in stations 1 and 2.

Moments. We now determine (higher) moments for the case $n=2$. We have that $\mathbb{E}X_1$, $\mathbb{E}X_1(X_1-1)$ and $\mathrm{Var}(X_1)$ are given in Equations (24)–(26). Furthermore, $\mathbb{E}X_2$, $\mathbb{E}X_1X_2$ and $\mathrm{cov}(X_1,X_2)$ are given in Equations (10), (13) and (14), and can also be obtained from Equation (27) by differentiation. In particular,

$$\mathrm{cov}(X_1,X_2)=-\frac{\lambda^2\mu_1}{(\mu_1+\tau_1(1-a_1)+\mu_2+\tau_2(1-a_2))(\mu_1+\tau_1(1-a_1))^2}. \quad (34)$$

It turns out that, not surprising, X_1 and X_2 are negatively correlated.
Twice differentiating (28), we see that (see also Equation (15))

$$[\mu_2+\tau_2(1-a_2^2)]\mathbb{E}X_2(X_2-1)=\mu_1\mathbb{E}X_1(X_1-1)+2\mu_1\mathbb{E}X_1X_2, \quad (35)$$

and hence

$$\mathrm{Var}(X_2)=\frac{1}{\mu_2+\tau_2(1-a_2^2)}[\frac{2\lambda^2\mu_1}{(\mu_1+\tau_1(1-a_1))(\mu_1+\tau_1(1-a_1^2))} \quad (36)$$
$$+\frac{2\lambda\mu_1}{\mu_1+\tau_1(1-a_1)+\mu_2+\tau_2(1-a_2)}\frac{\lambda}{\mu_1+\tau_1(1-a_1)}\frac{\mu_1}{\mu_2+\tau_2(1-a_2)}]$$
$$+\frac{\lambda\mu_1}{(\mu_1+\tau_1(1-a_1))(\mu_2+\tau_2(1-a_2))}-(\frac{\lambda\mu_1}{(\mu_1+\tau_1(1-a_1))(\mu_2+\tau_2(1-a_2))})^2.$$

Differentiating Equation (30) twice w.r.t. z_2 and taking $z_2=1$ gives

$$[\mu_2+\tau_2(1-a_2^2)]\mathbb{E}X_2^{app}(X_2^{app}-1)=\mu_1\mathbb{E}X_1(X_1-1)+2\mu_1\mathbb{E}X_1\mathbb{E}X_2, \quad (37)$$

and thus the difference between $\mathbb{E}X_2(X_2-1)$ and its approximation equals $\frac{2\mu_1}{\mu_2+\tau_2(1-a_2^2)}\mathrm{cov}(X_1,X_2)$, a difference which disappears when X_1 and X_2 are independent.

Remark 4. *It is important to observe that the approximation for* $\mathbb{E}X_2$ *is* exact *in the three approximation approaches. This follows from the fact that* $\mathbb{E}X_1$ *is exact, combined with the fact that the flow balance equations* (9) *remain valid under the batch-size independence (or RMF) assumption. Another way to see it is to observe that replacing* $P(z_1, z_2)$ *in* (27) *by* $P(z_1, 1)P(1, z_2)$ *(Approximation 2) leads after differentiation to exactly the same relation between* $\mathbb{E}X_1$ *and* $\mathbb{E}X_2^{app}$ *as the one between* $\mathbb{E}X_1$ *and* $\mathbb{E}X_2$*.*

Remark 5. *The previous remark, in combination with* (35) *and* (37)*, also implies that*

$$\mathbb{E}[X_2^2] - \mathbb{E}[(X_2^{app})^2] = \mathrm{Var}(X_2) - \mathrm{Var}(X_2^{app}) = \frac{2\mu_1}{\mu_2 + \tau_2(1-a_2^2)} cov(X_1, X_2). \tag{38}$$

Quality of the approximation. We now consider some parameter choices to investigate the quality of the approximation $P^{app}(1, z)$.
Case 1: $\mu_1 \downarrow 0$**.** The above formulas reveal that the covariance of X_1 and X_2 tends to zero as μ_1. The variance of X_1 tends to a positive limit, whereas the variance of X_2 tends to zero as μ_1. Hence, the correlation coefficient tends to zero as $\sqrt{\mu_1}$. Long time periods between openings of gate 1 enable station 1 to approach steady state, and then the numbers of particles in both stations indeed become independent.
Case 2: $\mu_1 \to \infty$**.** Now the covariance tends to zero as $1/\mu_1^2$, whereas the variance of X_1 tends to zero as $1/\mu_1$ and the variance of X_2 tends to a constant. Hence, the correlation coefficient tends to zero as $1/\mu_1^{3/2}$.
Case 3: $\lambda \downarrow 0$**.** The covariance tends to zero as λ^2, whereas both variances tend to zero as λ; hence, the correlation coefficient tends to zero as λ.
Case 4: $\lambda \uparrow \infty$**.** $\mathrm{cov}(X_1, X_2)$, $\mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$ all grow like λ^2. Hence, the correlation coefficient of X_1 and X_2 tends to some fixed negative constant as $\lambda \uparrow \infty$.
Case 5: $\tau_1 = \tau_2 = \tau \to \infty$**.** The above formulas reveal for this case of fast consumption that $\mathrm{cov}(X_1, X_2)$ tends to zero as $1/\tau^3$, while $\mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$ tend to zero as $1/\tau$ and $1/\tau^2$, respectively. Hence, the correlation coefficient tends to zero as $1/\tau^{3/2}$.
Case 6: $\tau_1 = \tau_2 = 0$**.** This is a classic ASIP case without consumption. For $\mu_1 = \mu_2 = \mu$ and $\rho = \lambda/\mu$ we get for the correlation: $\frac{-\rho/2}{\sqrt{(\rho+1)(2\rho+1)}}$, which decreases from zero (for $\rho = 0$) to $-1/\sqrt{8}$ (for $\rho \to \infty$); this is in agreement with [32].
Case 7: $\mu_2 \to \infty$**.** The covariance and the variance of X_2 tend to zero as $1/\mu_2$, and the variance of X_1 is not influenced by μ_2. Hence, the correlation coefficient tends to zero as $1/\sqrt{\mu_2}$.
Case 8: $\mu_2 = 0$**.** There are various non-zero limits.

Finally, we observe that in Cases 1, 3, 5 and 7 (i.e., cases in which the correlation coefficient tends to zero), we have that $\mathrm{Var}(X_2^{app})$ tends to zero as fast as $\mathrm{Var}(X_2)$. In Case 2 (again, the correlation coefficient tends to zero), the exact and approximate variance of X_2 tend to exactly the same limit as $\mu_1 \to \infty$.

Correlation between X_1 *and* X_2 *in homogeneous ASIP.* Previously, a number of explicit results for expressions within the ASIP system were derived in several publications, namely [10], [32], and [38]. However, due to the computational complexity involved in evaluating these expressions, the scope of analysis was largely restricted to the homogeneous ASIP model. The homogeneous ASIP system has been shown to be optimal in terms of various efficiency measures, as demonstrated in [32] and [37]. In a homogeneous ASIP system, all gates have exponentially distributed inter-opening times with the same rate μ. We extend our investigation of the correlation between X_1 and X_2 to a homogeneous ASIP with consumption such that all gates have the same opening rate $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, and the same consumption rate $\tau_1 = \tau_2 = \cdots = \tau_n = \tau$ and $a_1 = a_2 = \cdots = a_n = a$. Under these conditions, $\mathrm{Var}(X_1), \mathrm{Var}(X_2)$ and $\mathrm{cov}(X_1, X_2)$ become

$$\mathrm{Var}(X_1) = \frac{\lambda}{\mu + \tau(1-a)}\Big[\frac{\lambda\mu + \lambda\tau(1-a)^2}{(\mu + \tau(1-a))(\mu + \tau(1-a^2))} + 1\Big], \tag{39}$$

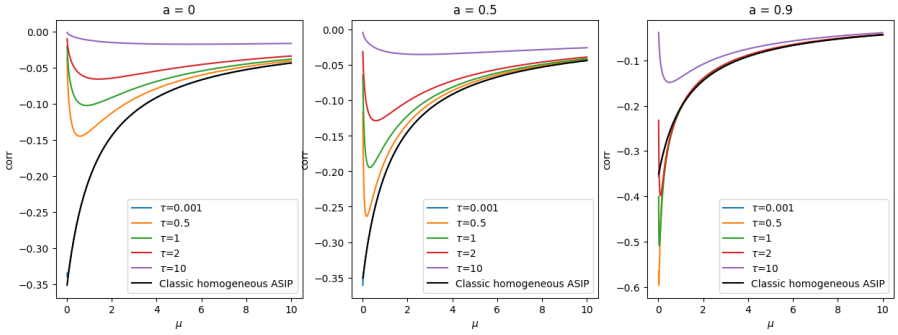

Figure 1: $\mathrm{Corr}(X_1, X_2)$ as a function of μ for the parameters τ and a.

$$\mathrm{Var}(X_2) = \frac{1}{\mu+\tau(1-a^2)}\left[\frac{2\lambda^2\mu}{(\mu+\tau(1-a))(\mu+\tau(1-a^2))} + \frac{2\lambda^2\mu^2}{2(\mu+\tau(1-a))^3}\right] + \frac{\lambda\mu}{(\mu+\tau(1-a))^2} - \frac{\lambda^2\mu^2}{(\mu+\tau(1-a))^4}, \quad (40)$$

and respectively

$$\mathrm{cov}(X_1, X_2) = -\frac{\lambda^2\mu}{2(\mu+\tau(1-a))^3}. \quad (41)$$

Using (39), (40) and (41), the correlation between the first two sites is found to be

$$\mathrm{Corr}(X_1, X_2) = \frac{-\frac{\lambda\mu}{2(\mu+\tau(1-a))^2}}{\sqrt{\left[\frac{\lambda\mu+\lambda\tau(1-a)^2}{(\mu+\tau(1-a))(\mu+\tau(1-a^2))}+1\right]\left[\frac{2\lambda\mu}{(\mu+\tau(1-a^2))^2}+\frac{\lambda\mu^2}{(\mu+\tau(1-a^2))(\mu+\tau(1-a))^2}+\frac{\mu}{\mu+\tau(1-a)}-\frac{\lambda\mu^2}{(\mu+\tau(1-a))^3}\right]}}. \quad (42)$$

In Figure 1, the comparison between the correlation of X_1 and X_2 as a function of μ is presented for the parameters τ and a in a homogeneous ASIP with and without consumption. The correlation between the two sites is depicted for a fixed value of $\lambda = 1$ and compared to the correlation in the classic homogeneous ASIP without consumption. When the consumption rate is low, indicated by small values of τ, the correlation between the sites converges rapidly towards the correlation observed in the classic ASIP without consumption. Moreover, for small values of μ, the correlation tends to overestimate its equivalent value in the classic ASIP. However, when the value of τ is large compared to μ, indicating a high consumption rate relative to the rate of gate openings, the correlation between the two sites tends to approach zero. This observation is understandable because during periods of high consumption rates, a substantial number of particles being consumed within their respective sites leads to a reduction in the correlation between them.

3.3. The case of general n

Taking $z_j = 1$ for all j unequal to $i-1$ and i in (3) (with $i \geq 3$), and assuming that the number of customers at station $i-2$ is independent of the numbers at the two stations $i-1$ and i, we obtain the following equation for the PGF $R(z_{i-1}, z_i)$ of the steady-state joint distribution of numbers of particles (X_{i-1}, X_i) in stations $i-1$ and i:

$$[\mu_{i-2}(1-\mathbb{E}z_{i-1}^{X_{i-2}}) + \tau_{i-1} + \tau_i + \mu_{i-1} + \mu_i]R(z_{i-1}, z_i) = \tau_{i-1}R(a_{i-1}z_{i-1} + 1 - a_{i-1}, z_i) + \tau_i R(z_{i-1}, a_i z_i + 1 - a_i) + \mu_{i-1}R(z_i, z_i) + \mu_i R(z_{i-1}, 1). \quad (43)$$

This equation has the same structure as (27) for the two-station tandem queue, but with the term $\lambda(1-z_1)$ in the left-hand side being replaced by $\mu_{i-2}(1-\mathbb{E}z_{i-1}^{X_{i-2}})$, representing batch arrivals from station $i-2$. If subsequently we focus on station i and once again employ the independence assumption (now between the numbers at stations $i-1$ and i), we get a recursion of exactly the same form as Equations (19) and (30): with $R_{i-1}(\cdot)$ the already obtained approximate PGF of the number of particles in station $i-1$, and $R_i(\cdot)$ the PGF of the number of particles in station i, the independence assumption yields

$$[\mu_{i-1}(1-R_{i-1}(z_i))+\tau_i+\mu_i]R_i(z_i)=\tau_i R_i(a_i z_i+1-a_i)+\mu_i. \tag{44}$$

This implies that successive applications of the independence assumption each time result to the same recursion, but with different parameters and the $R_{i-1}(z_i)$ obtained in the preceding step. The solution of such a recursion is already provided in Equation (31). Last, we observe that this approximation gives the *exact* mean queue length at each queue, as can be seen using either of the two reasonings in Remark 4.

4. Individual consumption

In this section, we consider the general model, again with one exception: now all τ_i are assumed to be zero (no binomial consumption). In Section 4.1, we determine the PGF of the number of particles in the first station. Section 4.2 discusses the model with two stations in series, applying the same approximation ideas as in the previous section for the PGF of the steady-state joint distribution of particles over the two stations. They again lead to the same approximation. In Section 4.3, we show how this approximation can be extended to the n-station case.

4.1. The case $n=1$

In this section, we determine the *distribution* of the number of particles in station 1 in steady state. We do this in two different ways. First, we use (3) to conclude that $P(z)\equiv P(z,1,\ldots,1)$ satisfies the following first-order inhomogeneous differential equation (this equation was already studied by Jenneskens [25]):

$$\frac{\mathrm{d}}{\mathrm{d}z}P(z)=(\frac{\lambda}{\nu_1}+\frac{\mu_1}{\nu_1(1-z)})P(z)-\frac{\mu_1}{\nu_1(1-z)}. \tag{45}$$

This equation corresponds to the equation for the PGF of the steady-state number of customers in the $M/M/\infty$ model with only total catastrophes. This model is a special case of the immigration/birth-death process with total catastrophes considered in Chao and Zheng [13]. Formulas (35) and (56) in [13] correspond to our formulas (47) and (48) below.

A straightforward application of the variation of constants method, with boundary condition $P(1)=1$, yields:

$$P(z)=\mathrm{e}^{-\frac{\lambda}{\nu_1}(1-z)}\int_z^1\frac{\mu_1}{\nu_1}\frac{(1-t)^{\frac{\mu_1}{\nu_1}-1}}{(1-z)^{\frac{\mu_1}{\nu_1}}}\mathrm{e}^{\frac{\lambda}{\nu_1}(1-t)}\mathrm{d}t. \tag{46}$$

Using the substitution $y=\frac{1-t}{1-z}$ we rewrite this into

$$P(z)=\int_0^1\frac{\mu_1}{\nu_1}y^{\frac{\mu_1}{\nu_1}-1}\mathrm{e}^{-\frac{\lambda}{\nu_1}(1-y)(1-z)}\mathrm{d}y. \tag{47}$$

Recognizing that $\mathrm{e}^{-\gamma(1-z)}$ is the PGF of a Poisson(γ) distributed random variable, we can invert to obtain the steady-state queue length distribution:

$$p(j)=\int_0^1\frac{\mu_1}{\nu_1}y^{\frac{\mu_1}{\nu_1}-1}\frac{(\frac{\lambda}{\nu_1}(1-y))^j}{j!}\mathrm{e}^{-\frac{\lambda}{\nu_1}(1-y)}\mathrm{d}y,\quad j=0,1,\ldots \tag{48}$$

We now provide a completely different derivation of this theorem. It is based on the observation that, by the PASTA principle, $p(j)$ also equals the distribution of the number of particles in station 1 immediately before a gate opening. Considering one (regenerative) interval between two successive gate openings of station 1, this station behaves as an $M/M/\infty$ queue with arrival rate λ and service rate ν_1, starting empty. It is well known, and readily seen, that the number of customers $X(t)$ in an initially empty $M/G/\infty$ queue with service time distribution $G(\cdot)$ (for which we shall later take the $\exp(\nu_1)$ distribution) is Poisson distributed with rate $\lambda t \int_0^t (1-G(t-u))\frac{\mathrm{d}u}{t} = \lambda \int_0^t (1-G(u))\mathrm{d}u$. This can, e.g., be understood by realizing that the initial Poisson arrival process is *thinned*, with thinning probability $\int_0^t (1-G(t-u))\frac{\mathrm{d}u}{t}$. Put differently, if there are k arrivals at station 1 in $[0,t]$, then these arrivals are uniformly distributed over $[0,t]$; and an arbitrary arrival then is still present with the above thinning probability.

Now consider the number of customers $X(T)$ at the time $T \sim \exp(\mu_1)$ of the first gate opening of station 1. Then

$$p(j) = \mathbb{P}(X(T)=j) = \int_0^\infty \mu_1 \mathrm{e}^{-\mu_1 t} \frac{(\lambda \int_0^t (1-G(u))\mathrm{d}u)^j}{j!} \mathrm{e}^{-\lambda \int_0^t (1-G(u))\mathrm{d}u} \mathrm{d}t. \tag{49}$$

Restricting ourselves to $1-G(t) = \mathrm{e}^{-\nu_1 t}$, and introducing the transform $y := \mathrm{e}^{-\nu_1 t}$, we readily retrieve (48).

Using that the mean of a Poisson(γ) distribution equals γ, it follows that the mean number of customers (for the case of general $G(\cdot)$) equals

$$\mathbb{E}X(T) = \sum_{j=0}^\infty j p(j) = \int_0^\infty \mu_1 \mathrm{e}^{-\mu_1 t} \lambda \int_0^t (1-G(u))\mathrm{d}u\mathrm{d}t = \lambda \int_0^\infty (1-G(u))\mathrm{e}^{-\mu_1 u}\mathrm{d}u. \tag{50}$$

The last equality follows by swapping integrals. With G a generic service time and E a generic $\exp(\mu_1)$ distributed random variable, we can rewrite this mean as follows:

$$\mathbb{E}X(T) = \frac{\lambda}{\mu_1} \int_0^\infty (1-G(u))\mu_1 \mathrm{e}^{-\mu_1 u}\mathrm{d}u = \frac{\lambda}{\mu_1}\, \mathbb{P}(G>E). \tag{51}$$

This formula has an obvious interpretation: There are on average λ/μ_1 arrivals per gate interval, and an arbitrary arrival has probability $\mathbb{P}(G>E)$ to be still present when the gate opens. When $1-G(t) = \mathrm{e}^{-\nu_1 t}$, (51) reduces to (8) (with $\tau_1 = 0$).

Remark 6. *Let us restrict ourselves again to the exp(ν_1) service time distribution. Next to the mean $\mathbb{E}X(T)$ (which equals $\mathbb{E}X_1$ because of PASTA), one can also obtain the variance of $X(T)$ (which equals Var(X_1)):*

$$\mathrm{Var}(X(T)) = \mathrm{Var}X_1 = \frac{\lambda}{\mu_1+\nu_1} \frac{\lambda\mu_1 + (\mu_1+\nu_1)(\mu_1+2\nu_1)}{(\mu_1+\nu_1)(\mu_1+2\nu_1)}. \tag{52}$$

This result also follows from (12) by taking $\tau_1 = 0$. Notice that the variance is larger than the mean, whereas it would be equal to the mean for a Poisson distributed random variable; Formula (48) shows that $X(T)$ is randomized Poisson$(\frac{\lambda}{\nu_1}(1-Y))$ distributed, where $\mathbb{P}(Y<y) = y^{\frac{\mu_1}{\nu_1}}$ for $0 \le y \le 1$.

4.2. The case $n = 2$

In this section, we discuss the general model with two stations. It follows from (3) that the two-dimensional PGF of the steady-state joint distribution of numbers of particles (X_1, X_2) in both stations is given by

$$\begin{aligned}[\lambda(1-z_1)+\mu_1+\mu_2]P(z_1,z_2) &= \nu_1[(1-p_1)(1-z_1)+p_1(z_2-z_1)]\frac{\partial}{\partial z_1}P(z_1,z_2) \\ &\quad + \nu_2(1-z_2)\frac{\partial}{\partial z_2}P(z_1,z_2) + \mu_1 P(z_2,z_2) + \mu_2 P(z_1,1). \end{aligned} \tag{53}$$

The $P(z_2, z_2)$ term makes this partial differential equation prohibitively difficult to solve. That even seems to hold for the one-dimensional differential equation that arises by taking $z_1 = 1$:

$$[\mu_1+\mu_2]P(1,z_2) = \nu_1 p_1(z_2-1)\frac{\partial}{\partial z_1}P(z_1,z_2)|_{z_1=1} + \nu_2(1-z_2)\frac{\mathrm{d}}{\mathrm{d}z_2}P(1,z_2) + \mu_1 P(z_2,z_2) + \mu_2. \quad (54)$$

Below we apply the same approximation ideas as suggested in the previous section, to approximate the marginal PGF of X_2; again, they lead to the same equation. This time we firstly assume that X_2 is independent of X_1. This independence would hold when $\mu_1 = 0$; a model of two infinite-server queues in series (with a gate at the second one) results, for which it is known that the steady-state queue length in the second queue of infinite-server queues in series does not depend on the queue length of the first queue. When $\mu_1 = \infty$, the queue lengths are also independent, as then $X_1 \equiv 0$. At the end of the section, we investigate the accuracy of the independence assumption in much more detail.

Under the independence assumption we have $P(z_1, z_2) = P(z_1, 1)P(1, z_2)$, and hence, with $P(z_2) = P(z_2, 1)$ given in (47),

$$\frac{\mathrm{d}}{\mathrm{d}z_2}P(1,z_2) = \frac{\nu_1 p_1}{\nu_2}\mathbb{E}X_1 P(1,z_2) + [\frac{\mu_1(1-P(z_2))}{\nu_2(1-z_2)} + \frac{\mu_2}{\nu_2(1-z_2)}]P(1,z_2) - \frac{\mu_2}{\nu_2(1-z_2)}. \quad (55)$$

Solving this differential equation, with boundary condition $P(1, 1) = 1$, gives the following approximation for $P(1, z_2)$:

$$P^{app}(1,z_2) = \int_{z_2}^{1} \frac{\mu_2}{\nu_2}\frac{(1-t)^{\frac{\mu_2}{\nu_2}-1}}{(1-z_2)^{\frac{\mu_2}{\nu_2}}} \mathrm{e}^{\frac{\mu_1}{\nu_2}\int_t^{z_2}\frac{1-P(v)}{1-v}\mathrm{d}v} \mathrm{e}^{\frac{\nu_1 p_1}{\nu_2}\mathbb{E}X_1(z_2-t)}\mathrm{d}t. \quad (56)$$

Application of l'Hôpital confirms that this expression indeed satisfies $P(1, 1) = 1$.

For our second approximation approach, we observe the following. The arrival process at station 2 is a sum of individual arrivals and arrivals of batches; the latter have PGF $P(z) = P(z, 1)$ as given in Equation (47) and occur according to a Poisson process with rate μ_1. However, the size of a batch is correlated with the length of the preceding gate opening. It implies that we cannot apply the same solution procedure as for the 1-station case of the previous section: if there are k batch arrivals at station 2 between two successive gate openings of that station, then we need to take the interarrival times of those batches into account instead of just reasoning that each batch arrival is uniformly distributed in the interval between those two gate openings. The approximation that we propose again is to ignore the dependence between the size of a batch and the preceding interarrival time. In addition, we assume that individual arrivals to station 2 occur according to a Poisson process with rate $\nu_1 p_1 \mathbb{E}X_1 = \nu_1 p_1 \frac{\lambda}{\mu_1+\nu_1}$. It is easily seen that the balance equations for the steady-state probability $\pi(j)$ of having j particles in the resulting infinite server queue for all $j = 0, 1, \ldots$ are given by

$$[\nu_1 p_1 \mathbb{E}X_1 + \mu_1 + j\nu_2 + \mu_2 I(j>0)]\pi(j) = \nu_1 p_1 \mathbb{E}X_1 \pi(j-1) + \mu_1 \sum_{k=0}^{j-1}\pi(k)p(j-k)I(j>0) + (j+1)\nu_2\pi(j+1) + \mu_2\sum_{k=1}^{\infty}\pi(k)I(j=0), \quad (57)$$

with $p(j)$ the probability of a batch arrival having size j. With $\Pi(z)$ the PGF of the $\pi(j)$ and $P(z)$ the PGF of the $p(j)$ (as given in Equation (47)), we obtain that $\Pi(z)$ satisfies the following differential equation, with boundary condition $\Pi(1) = 1$:

$$\frac{\mathrm{d}}{\mathrm{d}z}\Pi(z) = \frac{\nu_1 p_1}{\nu_2}\mathbb{E}X_1\Pi(z) + [\frac{\mu_1(1-P(z))}{\nu_2(1-z)} + \frac{\mu_2}{\nu_2(1-z)}]\Pi(z) - \frac{\mu_2}{\nu_2(1-z)}. \quad (58)$$

Comparison with (55) reveals that this is *exactly the same differential equation*. In other words: both approximations 1 and 2 amount to the same. The second approximation apparently implies that the two numbers of particles become independent.

Moments. Below we determine higher moments for the case $n = 2$. We see that $\mathbb{E}X_1$ and $\mathbb{E}X_2$ are given in Equations (8) and (10) with $\tau_1 = \tau_2 = 0$ respectively, while $\mathrm{Var}(X_1)$ is given in Equation (12). Incidentally, just like in Remark 4, it can be readily seen that differentiating $P^{app}(1, z_2)$ gives the *exact* mean queue length at station 2.

Differentiating (53) w.r.t. both z_1 and z_2 and taking $z_1 = z_2 = 1$ yields

$$(\mu_1 + \mu_2 + \nu_1 + \nu_2)\mathbb{E}X_1X_2 = \lambda\mathbb{E}X_2 + \nu_1 p_1 \mathbb{E}X_1(X_1 - 1), \tag{59}$$

and hence

$$\mathbb{E}X_1X_2 = \frac{\lambda^2}{\mu_1 + \nu_1 + \mu_2 + \nu_2}\left[\frac{1}{\mu_2 + \nu_2}\frac{\mu_1 + \nu_1 p_1}{\mu_1 + \nu_1} + \frac{2\nu_1 p_1}{(\mu_1 + \nu_1)(\mu_1 + 2\nu_1)}\right], \tag{60}$$

and finally

$$\begin{aligned}\mathrm{cov}(X_1, X_2) &= \frac{\lambda^2}{(\mu_1 + \nu_1 + \mu_2 + \nu_2)(\mu_1 + \nu_1)}[\frac{\mu_1 + \nu_1 p_1}{\mu_2 + \nu_2} + \frac{2\nu_1 p_1}{\mu_1 + 2\nu_1}] \\ &\quad - \frac{\lambda^2}{(\mu_1 + \nu_1)(\mu_2 + \nu_2)}\frac{\mu_1 + \nu_1 p_1}{\mu_1 + \nu_1} \\ &= -\frac{\lambda^2\mu_1}{(\mu_1 + \nu_1 + \mu_2 + \nu_2)(\mu_1 + \nu_1)^2}\frac{\mu_1 + 2\nu_1 - \nu_1 p_1}{\mu_1 + 2\nu_1}.\end{aligned} \tag{61}$$

This covariance, which could also have been derived from (14) by taking $\tau_1 = \tau_2 = 0$, is non-positive, and becomes zero if $\mu_1 = 0$. The negative correlation makes sense, because a relatively large number of particles in station 1 typically suggests that the last gate opening from station 1 occurred relatively long ago, leading to relatively few particles in station 2. The zero correlation when furthermore $\mu_1 = 0$ also makes sense, because $\mu_1 = 0$ corresponds to the first station being an ordinary infinite-server queue without gate openings.

Differentiating (54) twice w.r.t. z_2 and taking $z_2 = 1$ yields

$$(\mu_2 + 2\nu_2)\mathbb{E}X_2(X_2 - 1) = \mu_1\mathbb{E}X_1(X_1 - 1) + 2(\mu_1 + \nu_1 p_1)\mathbb{E}X_1X_2, \tag{62}$$

and after some further calculations we obtain

$$\begin{aligned}\mathrm{Var}(X_2) &= \frac{2\lambda^2}{(\mu_2 + 2\nu_2)(\mu_1 + \nu_1)}[\frac{\mu_1}{\mu_1 + 2\nu_1} \\ &\quad + \frac{(\mu_1 + \nu_1 p_1)^2}{(\mu_2 + \nu_2)(\mu_1 + \mu_2 + \nu_1 + \nu_2)} + \frac{2\nu_1 p_1(\mu_1 + \nu_1 p_1)}{(\mu_1 + 2\nu_1)(\mu_1 + \mu_2 + \nu_1 + \nu_2)}] \\ &\quad + \frac{\lambda(\mu_1 + \nu_1 p_1)}{(\mu_1 + \nu_1)(\mu_2 + \nu_2)} - (\frac{\lambda(\mu_1 + \nu_1 p_1)}{(\mu_1 + \nu_1)(\mu_2 + \nu_2)})^2.\end{aligned} \tag{63}$$

Let us now turn once more to Approximation 1 above. Differentiating (55) twice w.r.t. z_2 and taking $z_2 = 1$ gives

$$(\mu_2 + 2\nu_2)\mathbb{E}X_2^{app}(X_2^{app} - 1) = \mu_1\mathbb{E}X_1(X_1 - 1) + 2(\mu_1 + \nu_1 p_1)\mathbb{E}X_1\mathbb{E}X_2. \tag{64}$$

The difference between the exact result and the approximation, in the righthand sides of (62) and (64), is a factor (here $2(\mu_1 + \nu_1 p_1)\ \mathrm{cov}(X_1, X_2)$) which disappears when X_1 and X_2 are independent. Because $\mathbb{E}X_2 = \mathbb{E}X_2^{app}$, we have

$$\mathbb{E}[X_2^2] - \mathbb{E}[(X_2^{app})^2] = \mathrm{Var}(X_2) - \mathrm{Var}(X_2^{app}) = \frac{2(\mu_1 + \nu_1 p_1)}{\mu_2 + 2\nu_2}\mathrm{cov}(X_1, X_2). \tag{65}$$

Quality of the approximation. We consider some parameter choices to investigate the quality of the approximation $Q^{app}(1, z_2)$.

Case 1: $\mu_1 \downarrow 0$**.** The above formulas reveal that the covariance of X_1 and X_2 tends to zero as

μ_1. The variances of X_1 and X_2 (the latter as long as $p_1 > 0$) tend to positive limits. Hence, the correlation coefficient tends to zero as μ_1. Long openings of gate 1 enable station 1 to approach steady state, and then the numbers of particles in both stations indeed become independent. Notice that $\mu_1 = \mu_2 = 0$ corresponds to the classic $M/M/\infty - ./M/\infty$ tandem queue (in particular when $p_1 = 1$), in which the numbers of customers are independent.

Cases 2: $\mu_1 \to \infty$**, 3:** $\lambda \downarrow 0$**, 4:** $\lambda \uparrow \infty$**, 7:** $\mu_2 \to \infty$**.** The conclusions are exactly the same as for Cases 2, 3, 4, and 7 in Section 3.2.

Case 5: $\nu_1 = \nu_2 = \nu \to \infty$**.** The above formulas reveal for this case of fast consumption that $\text{cov}(X_1, X_2)$ tends to zero as $1/\nu^3$, while $\text{Var}(X_1)$ and $\text{Var}(X_2)$ both tend to zero as $1/\nu$. Hence, the correlation coefficient tends to zero as $1/\nu^2$.

Case 6: $\nu_1 = \nu_2 = 0$. This is a classic ASIP case without consumption. For $\mu_1 = \mu_2 = \mu$ and $\rho = \lambda/\mu$ we get for the correlation: $\frac{-\rho/2}{\sqrt{(\rho+1)(2\rho+1)}}$, which decreases from zero (for $\rho = 0$) to $-1/\sqrt{8}$ (for $\rho \to \infty$); this is in agreement with [32].

Case 8: $\mu_2 = 0$**.** There are various non-zero limits.

Finally, we observe that in Cases 3, 5 and 7 (i.e., cases in which the correlation coefficient tends to zero) $\text{Var}(X_2^{app})$ tends to zero as fast as $\text{Var}(X_2)$. In Cases 1 and 2 (again, the correlation coefficient tends to zero), the exact and approximate variance of X_2 tend to exactly the same limit as $\mu_1 \downarrow 0$, respectively $\mu_1 \to \infty$.

Correlation and variance approximation in a homogeneous ASIP. As in the previous section, we extend our analysis of the sites correlation and the variance approximation to a homogeneous ASIP with consumption such that all gates have the same opening rate $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, and the same consumption rate $\nu_1 = \nu_2 = \cdots = \nu_n = \nu$ and $p_1 = p_2 = \cdots = p_n = p$. Under these conditions, $\text{Var}(X_1), \text{Var}(X_2)$ and $\text{cov}(X_1, X_2)$ become

$$\text{Var}(X_1) = \frac{\lambda}{\mu+\nu}\left(\frac{\lambda\mu}{(\mu+\nu)(\mu+2\nu)} + 1\right), \tag{66}$$

$$\text{Var}(X_2) = \frac{2\lambda^2}{(\mu+2\nu)(\mu+\nu)}\left[\frac{\mu}{\mu+2\nu} + \frac{(\mu+\nu p)^2}{2(\mu+\nu)^2} + \frac{\nu p(\mu+\nu p)}{(\mu+2\nu)(\mu+\nu)}\right] + \frac{\lambda(\mu+\nu p)}{(\mu+\nu)^2} - \frac{\lambda^2(\mu+\nu p)^2}{(\mu+\nu)^4}, \tag{67}$$

and respectively

$$\text{cov}(X_1, X_2) = -\frac{\lambda^2\mu}{2(\mu+\nu)^3}\frac{\mu+2\nu-\nu p}{\mu+2\nu}. \tag{68}$$

The exact expression of $\text{Var}(X_2)$ for a homogeneous ASIP system with consumption (which is obtained in Equation (67)) and its approximation (which can be derived from Equation (64)) are compared in Figure 2. The comparisons between $\text{Var}(X_2)$ and its approximation are illustrated as a function of μ for the set of parameters ν and p and for $\lambda = 1$. Note that for the different values of ν and p when the value of μ is small compared to λ, the approximation tends to overestimate the true value of $\text{Var}(X_2)$, otherwise although the assumption of independence between sites is not accurate, it still yields a reasonably accurate approximation for the variance.

Using Equations (66), (67) and (68), the correlation between the first two sites is found to be

$$\text{Corr}(X_1, X_2) = \frac{-\frac{\lambda\mu}{2(\mu+\nu)^2}\left(\frac{\mu+2\nu-\nu p}{\mu+2\nu}\right)}{\sqrt{\left(\frac{\lambda\mu}{(\mu+\nu)(\mu+2\nu)}+1\right)\left(\frac{2\lambda}{\mu+2\nu}\left[\frac{\mu}{\mu+2\nu} + \frac{(\mu+\nu p)^2}{2(\mu+\nu)^2} + \frac{\nu p(\mu+\nu p)}{(\mu+2\nu)(\mu+\nu)}\right] + \frac{\mu+\nu p}{\mu+\nu} - \frac{\lambda(\mu+\nu p)^2}{(\mu+\nu)^3}\right)}}. \tag{69}$$

The correlation between X_1 and X_2 that is obtained in equation (69) for a homogeneous ASIP with consumption is evaluated and illustrated in Figure 3. This correlation is presented as a function of μ for the set of parameters ν and p and for $\lambda = 1$ and compared to the classic homogeneous ASIP (without consumption). Similar to the previous section, when considering

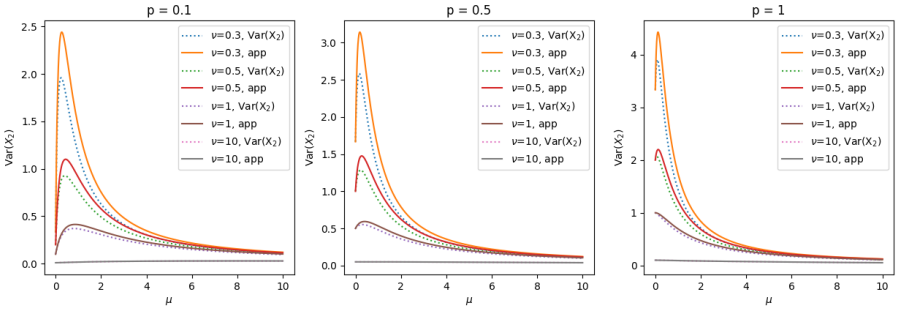

Figure 2: $Var(X_2)$ compared to its approximation.

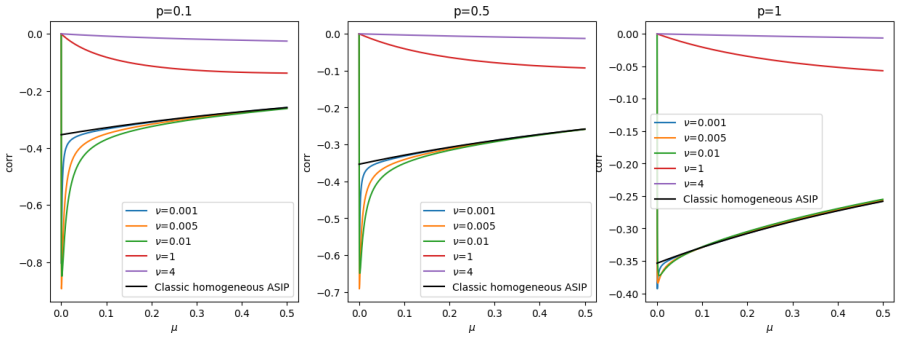

Figure 3: Corr(X_1, X_2) as a function of μ for the parameters ν and p.

small values of ν, indicating a low consumption rate, the correlation between the sites converges rapidly to the correlation observed in the classic ASIP without consumption. This behavior is reasonable since a decrease in the consumption rate causes the system to resemble more closely an ASIP without consumption. It is worth noting that for small values of μ, the correlation tends to overestimate its equivalent value in the classic ASIP. However, for large values of ν, representing a high consumption rate, the correlation between the two sites approaches zero. This observation is plausible because during periods of high consumption rates, the substantial amount of particles being consumed within their respective sites reduces the correlation between them.

4.3. The case of general n

Taking $z_j = 1$ for all j unequal to $i-1$ and i in Equation (3) (with $i \geq 3$), and assuming that the number of customers at station $i-2$ is independent of the numbers at the two stations $i-1$ and i, we obtain the following equation for the PGF $R(z_{i-1}, z_i)$ of the steady-state joint distribution of numbers of particles (X_{i-1}, X_i) in stations $i-1$ and i:

$$\begin{aligned}[\mu_{i-2}(1-\mathbb{E}z_{i-1}^{X_{i-2}}) + \nu_{i-2}p_{i-2}\mathbb{E}X_{i-2}(1-z_{i-1}) + \mu_{i-1} + \mu_i]R(z_{i-1}, z_i)\\ = \nu_{i-1}[(1-p_{i-1})(1-z_{i-1}) + p_{i-1}(z_i - z_{i-1})]\frac{\partial}{\partial z_{i-1}}R(z_{i-1}, z_i)\\ + \nu_i(1-z_i)\frac{\partial}{\partial z_i}R(z_{i-1}, z_i) + \mu_{i-1}R(z_i, z_i) + \mu_i R(z_{i-1}, 1). \quad (70)\end{aligned}$$

This equation has the same structure as (53) for the two-station tandem queue, but with the term $\lambda(1-z_1)$ in the left-hand side being replaced by the sum of the μ_{i-2} and ν_{i-2} terms, which represent batch and individual arrivals from station $i-2$. If, subsequently, we focus on station i and once again employ the independence assumption (now between the numbers at stations $i-1$ and i), we get a differential equation of exactly the same form as (55) and (58): with $R_{i-1}(\cdot)$ the

already obtained approximate PGF of the number of particles in station $i-1$, and $R_i(\cdot)$ the PGF of the number of particles in station i, the independence assumption yields

$$\frac{\mathrm{d}}{\mathrm{d}z_i}R_i(z_i) = \frac{\nu_{i-1}p_{i-1}}{\nu_i}\mathbb{E}X_{i-1}R_i(z_i) + [\frac{\mu_{i-1}(1-R_{i-1}(z_i))}{\nu_i(1-z_i)} + \frac{\mu_i}{\nu_i(1-z_i)}]R_i(z_i) - \frac{\mu_i}{\nu_i(1-z_i)}. \quad (71)$$

This implies that successive applications of the independence assumption each time yield the same differential equation, with different parameters and the $R_{i-1}(z_i)$ obtained in the preceding step. Equation (71) has exactly the same structure as (55), and its solution reads (cf. (56)):

$$R_i(z_i) = \int_{z_i}^1 \frac{\mu_i}{\nu_i}\frac{(1-t)^{\frac{\mu_i}{\nu_i}-1}}{(1-z_i)^{\frac{\mu_i}{\nu_i}}}\mathrm{e}^{\frac{\mu_{i-1}}{\nu_i}\int_t^{z_i}\frac{1-R_{i-1}(v)}{1-v}\mathrm{d}v}\mathrm{e}^{\frac{\nu_{i-1}p_{i-1}}{\nu_i}\mathbb{E}X_{i-1}(z_i-t)}\mathrm{d}t. \quad (72)$$

5. Conclusions and suggestions for further research

In this paper, we have introduced a very general tandem queueing model that contains the infinite-server TJN and ASIP models as special cases. We have shown how to obtain (joint) moments (of any order) of numbers of customers at all stations, and we have presented an approximation for the queue length distributions at the various stations.

In future research, it would be interesting to study other performance measures, such as (i) the probability for a particle to be consumed and (ii) the time it takes a particle to be either consumed or to travel through all n stations. Next, we intend to study ASIP tandem queues with additional individual consumptions or movements that are modelled via *single* servers at each station (next to gates that allow transfer of all particles to the next station). Further, we will apply the Power-Series Algorithm (PSA) (see [5] for a survey), which is a analytic-numerical approach, to ASIP. Due to the curse of dimensionality that hampers PSA, the applicability of this method may be restricted to small systems, but yields accurate and fast approximations. We also want to explore other approximations for the queue length distribution in ASIP tandem queues, e.g. by fitting the first few moments (using that we can recursively obtain these moments from Equation (5) in Section 2.3). In some application areas, it may be more natural to let the input to the ASIP be a fluid flow of constant rate. Still allowing random gate openings at all stations and allowing consumption, it would then be interesting to study the buffer contents (workloads) of all stations. One such case was treated in [10, Section 3]: a two-queue ASIP with Lévy input processes (which contains the case of fluid input) and with a constant (hence non-proportional) consumption rate at the first station.

Finally, let us elaborate a bit more on one research direction that we consider to be particularly interesting. In Sections 3.3 and 4.3, we already briefly indicate that the approximation methods are applicable to $Q_2, \ldots, Q_n$ in an n-queue tandem ASIP with consumption ($n \geq 3$). Our approach can also be extended to an ASIP *feed-forward network* with consumption. In such a network, particles arrive at $Q_{11}, \ldots, Q_{M1}$ according to independent Poisson processes. This is layer 1. The gate of Q_{i1} opens at $\exp(\mu_{i1})$ intervals. Its content then moves as a batch to Q_{j2} of layer 2, with probability $p_{i1,j2}$. Similarly, gates of the queues in layer 2 open at exponentially distributed intervals, and their content moves as a batch to one of the queues of layer 3, etc. One could also allow movements from a layer k to a layer $k+m$ with $m > 1$; but we do not allow feedback to lower layers. Such feed-forward networks arise naturally in a host of applications, including the transport of macromolecules from cell to cell.

Again, one can get exact expressions for moments and correlations. Furthermore, in such networks RMF again coincides with the assumption that the size of a batch is independent of the corresponding gate opening interval. We conjecture that this independence assumption becomes more and more accurate when the number of stations per layer grows. Our approximation methods again yield the exact mean queue lengths; and we believe that there is a convincing intuitive explanation for our conjecture, along the following lines. When Q_{jk} in layer k receives batches, this still occurs according to a Poisson process. The size of a batch still depends on the corresponding gate opening interval, but if there are many queues in layer $k-1$, then the batch

can come from many different queues and the effect of the above-mentioned dependence becomes negligible. This is very similar to the famous Independence Assumption of Kleinrock [28, Section 3.4] for message-switching communication networks. In such networks, messages maintain their size while travelling through the network. Kleinrock ignored the (in his case obviously very strong) dependence, assuming that the corresponding service times in successive queues are all independent. His Independence Assumption results in a bad approximation for tandem queues, but the approximation becomes better and better, the larger or more complex the network becomes – just like in our case, the effect of dependence becomes negligible if a message can come from many different stations.

References

[1] F. Baccelli, M. Davydov and T. Taillefumier (2022). Replica-Mean-Field limits of fragmentation-interaction-aggregation processes. *J. Appl. Probab.* **59**, 38—59.

[2] F. Baccelli and T. Taillefumier (2019). Replica-mean-field limits for intensity-based neural networks. *J. Appl. Dynam. Syst.* **18**, 1756—1797.

[3] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery* **22**, 248–260.

[4] D.P. Bertsekas and R.G. Gallager (1992). *Data Networks*. Prentice Hall, 2nd ed.

[5] J.P.C. Blanc (1993). Performance analysis and optimization with the power-series algorithm. In: L. Donatiello and R. Nelson (eds.), *Performance Evaluation of Computer and Communication Systems* (LNCS Vol. 729, Springer) pp. 53-80.

[6] J.S. Bonifacino and B.S. Glick (2004). The mechanisms of vesicle budding and fusion. *Cell* **116**, 153—166.

[7] R.J. Boucherie and N.M. van Dijk (eds.) (2011). *Queueing Networks: A Fundamental Approach*. Springer.

[8] O.J. Boxma (1984). $M/G/\infty$ tandem queues. *Stochastic Processes and their Applications* **18**, 153-164.

[9] O. Boxma, O. Kella, and U. Yechiali (2016). An ASIP model with general gate opening intervals. *Queueing Systems* **84** (1), 1-20.

[10] O. Boxma, O. Kella, and U. Yechiali (2021). Workload distributions in ASIP queueing networks. *Queueing Systems* **97** (1-2), 81-100.

[11] P. D. Brewer, E. N. Habtemichael, I. Romenskaia, C. C. Mastick and C. F. Adelle (2016). Glut4 is sorted from a Rab10 GTPase-independent constitutive recycling pathway into a highly insulin-responsive Rab10 GTPase-dependent sequestration pathway after adipocyte differentiation. *Journal of Biological Chemistry* **291** (2), 773–789.

[12] R. Bundschuh (2002). Asymmetric exclusion process and extremal statistics of random sequences. Phys. Rev. E, **65**, 031911.

[13] X. Chao and Y. Zheng (2003). Transient analysis of immigration/birth–death processes with total catastrophes. *Probability in the Engineering and Informational Sciences* **17**, 83–106.

[14] H.Chen and D.D. Yao (2001). *Fundamentals of Queueing Networks*. Springer.

[15] J.W. Cohen (1982). *The Single Server Queue*. North-Holland Publ. Cy., 2nd ed.

[16] J.W. Cohen and O.J. Boxma (1983). *Boundary Value Problems in Queueing System Analysis*. North-Holland Publ. Cy.

[17] B. Derrida (1998). An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. Phys. Rep. **65**, 301.

[18] G. Fayolle, R. Iasnogorodski and V. Malyshev (1999). *Random Walks in the Quarter-Plane*. Springer.

[19] D. Fiems, M. Mandjes and B. Patch (2018). Networks of infinite-server queues with multiplicative transitions. *Performance Evaluation* **123-124**, 35-49.

[20] O. Golinelli and K. Mallick (2006). The asymmetric simple exclusion process: An integrable model for non-equilibrium statistical mechanics. J. Phys. A Math. General, **39**, 12679.

[21] S. Huang, L. M. Lifshitz, C. Jones, K. D. Bellve, O. Standley, S. Fonseca, S. Corvera, K. E. Fogarty and M. P. Czech (2007). Insulin stimulates membrane fusion and GLUT4 accumulation in clathrin coats on adipocyte plasma membranes. *Molecular and Cellular Biology* **27** (9), 3456-3469.

[22] J.R. Jackson (1957). Networks of waiting lines. Oper. Res. **5**, 518–521.

[23] J.R. Jackson (1963). Jobshop-like queueing systems. *Management Science* **10**, 131-142.

[24] R.R.P. Jackson (1954). Queueing systems with phase type service. *Operational Research Quarterly* **5**, 109-120.

[25] T. Jenneskens (2018). *Cellular Transport: A Queueing Systems Analysis*. Bachelor Thesis, Eindhoven University of Technology.

[26] S. Kapodistria, T. Phung-Duc and J. Resing (2016). Linear birth/immigration death process with binomial catastrophes. *Probability in the Engineering and Informational Sciences* **30**, 79-111.

[27] F.P. Kelly (1979). *Reversibility and Stochastic Networks*. Wiley.

[28] L. Kleinrock (1964). *Communication Nets – Stochastic Message Flow and Delay*. Dover Publ., New York.

[29] V. A. Lizunov, K. Stenkula, A. Troy, S. W. Cushman and J. Zimmerberg (2013). Insulin regulates Glut4 confinement in plasma membrane clusters in adipose cells. *PLoS ONE* **8** (3), e57559.

[30] Y. Luyan and T. O. Taillefumier (2022). Metastable spiking networks in the replica mean-field limit. *PLoS Comput Biol* **18** (6): e1010215.

[31] S. Reuveni, I. Eliazar, O. Hirschberg, and U. Yechiali (2014). Occupation probabilities and fluctuations in the asymmetric inclusion process. *Physical Review* **E89** (4), 042109.
[32] S. Reuveni, I. Eliazar and U. Yechiali (2011). Asymmetric inclusion process. *Physical Review* **E84**, 041101-1 - 041101-16.
[33] S. Reuveni, I. Eliazar, and U. Yechiali (2012). Asymmetric inclusion process as a showcase of complexity. *Physical Review Letters* **109** (2), 020603.
[34] S. Reuveni, I. Eliazar, and U. Yechiali (2012). Limit laws for the asymmetric inclusion process. *Physical Review* **E86** (6), 061133.
[35] L.B. Shaw, R.K. Zia and K.H. Lee (2003). Totally asymmetric exclusion process with extended objects: A model for protein synthesis. Phys. Rev. E, **68**, 021910.
[36] Y. K. Xu, K. D. Xu, J.Y Li, L. Q. Feng, D. Lang, and X. X. Zheng (2007). Bi-directional transport of GLUT4 vesicles near the plasma membrane of primary rat adipocytes. *Biochemical and Biophysical Research Communications* **359** (1), 121-128.
[37] Y. Yeger and U. Yechiali (2022). Performance measures in a generalized asymmetric simple inclusion process. *MDPI* **10** (4), 594.
[38] Y. Yeger and U. Yechiali (2022). Matrix approach for analyzing n-site generalized ASIP systems: PGF and site occupancy probabilities. *MDPI* **10** (23), 4624.