EURANDOM PREPRINT SERIES

2020-007

June 3, 2020

Workload distributions in ASIP queueing networks

O. Boxma, O. Kella, U. Yechiali
ISSN 1389-2355

Workload distributions in ASIP queueing networks

Onno Boxma* , Offer Kella† , and Uri Yechiali‡

June 3, 2020

Abstract

The workload of a generalized n-site Asymmetric Simple Inclusion Process (ASIP) is investigated. Three models are analyzed. The first model is a serial network for which the steady-state Laplace-Stieltjes transform (LST) of the total workload in the first k sites ($k \leq n$) just after gate openings and at arbitrary epochs is derived. The former (just after gate openings) turns out to be an LST of the sum of k independent random variables. The second model is a 2-site ASIP with leakage from the first queue. Gate openings occur at exponentially distributed intervals and the external input processes to the stations are two independent subordinator Lévy processes. The steady-state joint workload distribution right *after* gate openings, right *before* gate openings and at *arbitrary* epochs is derived. The third model is a shot-noise counterpart of the second model where the workload at the first queue behaves like a shot-noise process. The steady-state total amount of work just before a gate opening turns out to be a sum of two independent random variables.

Keywords: ASIP queueing networks, Lévy networks.

1 Introduction

A tandem stochastic network is a linear set of n sites (queues) denoted Q_1,Q_2,...,Q_n to which a random stream of particles (or work) flows. Every site consists of a buffer and a gate behind it that opens according to some stochastic process. Each site is characterized by some buffer capacity C_{site}, denoting the maximal number of particles (amount of work) that the buffer can hold and by C_{gate}, the maximal number of particles (work) that can pass through the gate when it opens. Particles (work) flow into the system, usually to the first site, and then move uni-directionally from one site to the next, until exiting the system. Three fundamental models, distinguished by their C_{site} and C_{gate} values, have been analyzed in the literature: The first is a Tandem Jackson Network (TJN), where particles flow into the first site according to a Poisson process. The site capacities are $C_{\text{site}} = \infty$ and $C_{\text{gate}} = 1$, while each gate opens independently every exponentially distributed period of time, allowing at most a single particle

*EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (o.j.boxma@tue.nl); research partly funded by an NWO TOP grant, Grant Number 613.001.352, and by the NWO Gravitation project NETWORKS, Grant Number 024.002.003

†Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel (offer.kella@huji.ac.il); supported in part by grant 1647/17 from the Israel Science Foundation and the Vigevani Chair in Statistics

‡Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (uriy@tauex.tau.ac.il)

(if any) to hop to the next site. The TJN [10] [11] is famous for its product-form solution for the steady-state joint distribution function of the queue occupancies. The second model is the Asymmetric Exclusion Process (ASEP), a fundamental model in non-equilibrium statistical physics [13] [9], where $C_{\text{site}} = 1$ and $C_{\text{gate}} \geq 1$. If the gate of Q_i opens while the buffer of Q_{i+1} is not empty, the particle in Q_i is blocked. The third setup is the recently introduced [14][15][17] Asymmetric Inclusion Process (ASIP), where both $C_{\text{site}} = \infty$ and $C_{\text{gate}} = \infty$. As such, the ASIP fills the missing link between the TJN and the ASEP. The major difference between the models is that in the ASIP, when the gate of Q_i opens, all particles (work) present there move simultaneously and instantaneously to the buffer of the next site, joining the cluster of particles (work) there to form a larger cluster, while in the TJN or ASEP at most one particle moves forward when the site's gate opens. The ASIP may be considered as an inclusion counterpart of the ASEP and as a batch-service counterpart of the TJN. It was shown in [15] that, in contrast to the TJN, the ASIP does not admit a product-form solution for its steady-state joint distribution function of the queue occupancies. However, it admits a product-form solution for the site loads. ASIP's limit laws were treated in [16] and [17]. It was shown in [17] that, in a symmetric ASIP, the asymptotic probability that site k is occupied is proportional to $1/\sqrt{k}$. Occupation probabilities were further studied in [18]. The ASIP has been generalized in [5] to the case of general gate opening intervals, where gate openings are determined by a Markov renewal process. The focus in [5] is on the steady-state joint distribution function of the number of particles in the various sites. A very recent study [3] analyzed occupancy correlations in the classical ASIP.

The current paper focuses on the analysis of workload in an ASIP network. Three models are considered. The first is a serial model for which the steady-state Laplace-Stieltjes transform (LST) of the total workload in the first k sites is derived just after gate openings and at arbitrary epochs. The second model is an ASIP model consisting of only two sites in series, each with its own gate and a leakage of a fixed rate from Q_1. Gate openings occur at exponentially distributed intervals and the external input processes to the two sites are non-decreasing Lévy processes. The steady-state joint workload distribution functions right after gate openings, right before gate openings and at arbitrary epochs are derived. The third model is a shot-noise counterpart of the second model where the leakage rate from the first queue is linear in the workload and thus, in between gate openings, behaves like a shot noise process. We obtain the steady-state joint workload LST just before and just after gate openings. Sections 2, 3 and 4 treat Models 1,2 and 3, respectively.

2 Model 1: n queues in series

This section is devoted to an ASIP model consisting of n queues in series. The model is described in Subsection 2.1. In Subsection 2.2 we derive an explicit expression for the steady-state Laplace-Stieltjes transform (LST) of the total workload in the first k queues, just after a gate opening. The workload LST in the first k queues at arbitrary epochs is derived in Subsection 2.3. The steady-state joint workload LST right after gate openings is harder to obtain. In Subsection 2.4 we provide a fairly detailed procedure for obtaining it in the cases $n = 2, 3$.

2.1 Model description

Consider the following model of n queues $Q_1, \ldots, Q_n$ in series. Each queue has one gate behind it, which may be viewed as a server. Gates are closed almost all the time. When gate $i = 1, 2, \ldots, n-1$ (the gate behind Q_i) opens, all the work present in Q_i is instantaneously transferred to Q_{i+1}. When gate n opens, all the work present in Q_n instantaneously leaves the system. After the transfer, the gate immediately closes again. Gate openings are determined by a Markov renewal process. If, at some time t, gate i opens, then with probability p_{ij} the next gate to open is gate j and the time until that gate opens is an independent random variable distributed like O_{ij}. We assume that the Markov chain governing the successive gate openings is irreducible and we denote its steady-state distribution by π_i, $i = 1, \ldots, n$.

During an O_{ij} period, work (sometimes denoted as fluid) may externally arrive at all queues. The LST of amounts of work arriving to $Q_1, \ldots, Q_n$ during an O_{ij} period is given by $A_{ij}(s_1, \ldots, s_n)$. Given O_{ij}, these amounts are independent of amounts arriving during previous periods. In addition, we denote the LST of the cumulative amount of work arriving to $Q_1, \ldots, Q_k$ during an O_{ij} period by $A_{ijk}(s) = A_{ij}(s, \ldots, s, 1, \ldots, 1)$, where the last s occurs at position k. Notice that one example is provided by an n-dimensional Lévy subordinator process, possibly with dependence between amounts arriving at different queues and with Laplace exponents which may depend on the type of gate opening interval.

We recall that we restrict ourselves to the case in which work from Q_i can only move to Q_{i+1}, $i = 1, 2, \ldots, n-1$. That assumption will allow us to obtain exact steady-state results for the total amount of work $V_{(k)}$ which is present in the first k queues right after a gate opening ($k = 1, 2, \ldots, n$). Our results will become somewhat simpler in the special case in which the next gate opening is of gate j with a fixed probability q_j, *i.e.,* irrespective of the index of the previous gate opening.

2.2 Analysis of the total workload in the first k queues

We are interested in the steady-state joint distribution of the amounts of work $(V_1, \ldots, V_n)$ just after a gate opening. To argue the existence of such a distribution, one can follow a similar reasoning as in Section 2 of [5], that also considers an ASIP model of n queues in series, but in which the focus is on customers instead of work/fluid.

In the present subsection we shall in particular focus on $V_{(k)} = V_1 + \cdots + V_k$, namely, the total amount of work in the first k queues right after a gate opening. It will turn out that the analysis of $V_{(k)}$ can closely follow the reasoning for queue lengths in [5].

Introducing M, the index of the gate that has just opened, we consider

$$\xi_{ki}(s) = \mathbb{E}[\mathrm{e}^{-sV_{(k)}} 1_{\{M=i\}}], \quad k, i = 1, \ldots, n, \tag{1}$$

where $1_{\{\cdot\}}$ denotes an indicator. The fact that fluid can only move to downstream queues (i.e., with higher index) will allow us to express all $\xi_{ki}(s)$ for a fixed k as functions of $\xi_{k-1,j}(s)$ and, inductively as functions of $\xi_{1j}(s)$, which can be determined explicitly.

Step 1: Determination of $\xi_{1j}(s)$, $j = 1, \ldots, n$.
Obviously

$$\xi_{11}(s) = \mathbb{P}(M = 1) = \pi_1 \,. \tag{2}$$

Indeed, after gate 1 has opened, Q_1 instantaneously has become empty. Now consider two successive gate openings in steady state, the latter one being an opening of gate j. Summing

over all possible gates i opened at the previous gate opening gives:

$$\xi_{1j}(s) = \sum_{i=1}^{n} \xi_{1i}(s) p_{ij} A_{ij1}(s) = \sum_{i=2}^{n} \xi_{1i}(s) p_{ij} A_{ij1}(s) + \xi_{11}(s) p_{1j} A_{1j1}(s), \quad j \neq 1. \tag{3}$$

Here we have employed $A_{ij1}(s)$, the LST of the amount of work arriving at Q_1 during the gate opening interval.

Introducing the $(n-1)$-dimensional vectors

$$\xi_1(s) = (\xi_{12}(s), \ldots, , \xi_{1n}(s)), \quad R_1(s) = (p_{12} A_{121}(s), \ldots, p_{1n} A_{1n1}(s)),$$

and with the matrix $P_1(s)$ of which the (i,j)th coordinate is $p_{ij} A_{ij1}(s)$, we can write (3) as

$$\xi_1(s) = \xi_1(s) P_1(s) + \xi_{11}(s) R_1(s), \tag{4}$$

and hence, with I the matrix with ones on the diagonal and zeroes outside the diagonal,

$$\xi_1(s) = \xi_{11}(s) R_1(s) (I - P_1(s))^{-1}. \tag{5}$$

All the terms in the righthand side of (5) are known; in particular, $\xi_{11}(s) = \pi_1$ is given in (2). Hence we have determined $\xi_{11}(s), \xi_{12}(s), \ldots, \xi_{1n}(s)$.

Step 2: Expressing $\xi_{kj}(s)$ in terms of $\xi_{k-1,i}(s)$, for $i, j = 1, \ldots, n$, $k = 2, \ldots, n$.
Considering two successive gate openings in steady state, the last one being of gate j, and summing over all possible gates i for the first gate opening, we have for $k = 2, \ldots, n$:

$$\xi_{kj}(s) = \sum_{i=1}^{n} \xi_{ki}(s) p_{ij} A_{ijk}(s) = \sum_{i \neq k} \xi_{ki}(s) p_{ij} A_{ijk}(s) + \xi_{kk}(s) p_{kj} A_{kjk}(s), \quad j \neq k, \tag{6}$$

whereas

$$\xi_{kk}(s) = \sum_{i=1}^{n} \xi_{k-1,i}(s) p_{ik} A_{ik,k-1}(s). \tag{7}$$

The explanation for the deviating terms ($\xi_{k-1,i}(s)$ instead of $\xi_{ki}(s)$ and $A_{ik,k-1}(s)$ instead of $A_{ikk}(s)$) is that Q_k has become empty right after an opening of gate k, so that the total amount of fluid present in $Q_1, \ldots, Q_k$ equals the total amount present in $Q_1, \ldots, Q_{k-1}$ after the previous gate opening, plus the amount of fluid arriving in the first $k-1$ queues.

Introducing the $(n-1)$-dimensional vectors

$$\xi_k(s) = (\xi_{k1}(s), \ldots, \xi_{k,k-1}(s), \xi_{k,k+1}(s), \ldots, \xi_{kn}(s)),$$

$$R_k(s) = (p_{k1} A_{k1k}(s), \ldots, p_{k,k-1} A_{k,k-1,k}(s), p_{k,k+1} A_{k,k+1,k}(s), \ldots, p_{kn} A_{knk}(s)),$$

and with the matrix $P_k(s)$ of which the (i,j)th coordinate is $p_{ij} A_{ijk}(s)$, we can write (6) as

$$\xi_k(s) = \xi_k(s) P_k(s) + \xi_{kk}(s) R_k(s), \tag{8}$$

yielding

$$\xi_k(s) = \xi_{kk}(s) R_k(s) (I - P_k(s))^{-1}. \tag{9}$$

Introducing
$C_{k-1}(s) = (p_{1k}A_{1k,k-1}(s), \ldots, p_{k-2,k}A_{k-2,k,k-1}(s), p_{kk}A_{kk,k-1}(s), \ldots, p_{nk}A_{nk,k-1}(s))$,
we can rewrite (7) as

$$\xi_{kk}(s) = \xi_{k-1}(s)C_{k-1}^T(s) + \xi_{k-1,k-1}(s)p_{k-1,k}A_{k-1,k,k-1}(s). \tag{10}$$

We have thus expressed $\xi_k(s)$ in terms of $\xi_{kk}(s)$ via (9), and $\xi_{kk}(s)$ in terms of $\xi_{k-1}(s)$ and $\xi_{k-1,k-1}(s)$ via (10). Iterating, defining an empty product to be one and defining $\xi_0(s)C_0^T(s)$ to equal π_1 for notational convenience, we obtain:

$$\xi_{kk}(s) = \sum_{i=0}^{k-1} \xi_i(s)C_i^T(s) \prod_{j=i+1}^{k-1} p_{j,j+1}A_{j,j+1,j}(s). \tag{11}$$

By carefully studying the structure of the above recursions, and introducing

$$H_i(s) = R_i(s)(I - P_i(s))^{-1}C_i^T(s), \quad i = 1, \ldots, n,$$

the following holds:

$$\xi_{kk}(s) = \pi_1 \sum_{\ell_1,\ldots,\ell_{k-1}\in\{0,1\}} \prod_{i=1}^{k-1} (\ell_i H_i(s) + (1-\ell_i)p_{i,i+1}A_{i,i+1,i}(s)), \quad k = 1, \ldots, n\,. \tag{12}$$

With (12) and (9) we have a recipe for determining $\xi_{kj}(s)$ explicitly, for $k, j = 1, \ldots, n$.

Example. Let us consider the special case in which $p_{ij} = p_{1j}$, $\forall\ i, j$, and $A_{ijk}(s) = A_{1jk}(s)$, $\forall\ i, j, k$. Namely, the Markov renewal process that determines the gate openings and the intervals in between has a simple structure. Each time the next gate opening is of gate j with probability p_{1j}, and the interval length until the next opening also only depends on j. In this case we can obtain a simple expression for $\mathbb{E}[\mathrm{e}^{-sV_{(k)}}] = \sum_{j=1}^n \xi_{kj}(s)$. We have:

$$\xi_{11}(s) = \pi_1 = p_{11}, \tag{13}$$

and summing (3) over $j = 2, \ldots, n$:

$$\mathbb{E}[\mathrm{e}^{-sV_{(1)}}] = \sum_{j=1}^n \xi_{1j}(s) = p_{11} + \sum_{j=2}^n p_{1j}A_{1j1}(s)\mathbb{E}[\mathrm{e}^{-sV_{(1)}}], \tag{14}$$

yielding

$$\mathbb{E}[\mathrm{e}^{-sV_{(1)}}] = \frac{p_{11}}{1 - \sum_{j=2}^n p_{1j}A_{1j1}(s)}. \tag{15}$$

Furthermore, summing (6) over $j \neq k$ and using (7),

$$\mathbb{E}[\mathrm{e}^{-sV_{(k)}}] = p_{1k}A_{1k,k-1}(s)\mathbb{E}[\mathrm{e}^{-sV_{(k-1)}}] + \sum_{j\neq k} p_{1j}A_{1jk}(s)\mathbb{E}[\mathrm{e}^{-sV_{(k)}}], \tag{16}$$

leading to the following recursive expression of $\mathbb{E}[\mathrm{e}^{-sV_{(k)}}]$ in terms of $\mathbb{E}[\mathrm{e}^{-sV_{(k-1)}}]$:

$$\mathbb{E}[\mathrm{e}^{-sV_{(k)}}] = \frac{p_{1k}A_{1k,k-1}(s)}{1 - \sum_{j\neq k} p_{1j}A_{1jk}(s)}\mathbb{E}[\mathrm{e}^{-sV_{(k-1)}}]. \tag{17}$$

Via iteration we obtain:

$$\mathbb{E}[\mathrm{e}^{-sV_{(k)}}] = \prod_{i=1}^{k} \frac{p_{1i}A_{1i,i-1}(s)}{1-\sum_{j\neq i} p_{1j}A_{1jk}(s)}, \tag{18}$$

where $A_{110}(s) = 1$.
Formula (18) reveals a decomposition property. That is, the LST is a product of k terms, all of which are LST's of random variables, and this implies that $V_{(k)}$ can be represented as sum of k independent random variables.

2.3 The workload distribution at an arbitrary epoch

Armed with the LST's $\xi_{ki}(s)$ from the previous subsection, we shall now derive an expression for the steady-state LST $\chi_k(s)$ of the total workload in the first k queues *at an arbitrary epoch*. In order to do this, we need to further specify the arrival process. Indeed, it clearly makes a difference whether the amounts of work which arrive in the queues during a gate opening interval O_{ij} enter the system at the beginning of such an interval, or at the end, or according to some other stochastic process. In this subsection we shall assume that the external arrival process is an n-dimensional subordinator (hence a non-decreasing Lévy process) which may vary from one gate interval to another. Denote the Lévy input process during an O_{ij} period by $\{X_{ij}^{(1)}(t), \ldots, X_{ij}^{(n)}(t), t \geq 0\}$ and its Laplace exponent by $-\eta_{ij}(s_1, \ldots, s_n)$, *i.e.,*

$$\mathbb{E}[\mathrm{e}^{-s_1 X_{ij}^{(1)}(t) - \cdots - s_n X_{ij}^{(n)}(t)}] = \mathrm{e}^{-t\eta_{ij}(s_1,\ldots,s_n)}.$$

Hence $A_{ij}(s_1, \ldots, s_n)$ is the LST of O_{ij} with parameter $\eta_{ij}(s_1, \ldots, s_n)$:

$$\begin{aligned} A_{ij}(s_1, \ldots, s_n) &= \mathbb{E}[\mathrm{e}^{-s_1 X_{ij}^{(1)}(O_{ij}) - \cdots - s_n X_{ij}^{(n)}(O_{ij})}] \\ &= \mathbb{E}[\mathrm{e}^{-\eta_{ij}(s_1,\ldots,s_n)O_{ij}}]. \end{aligned} \tag{19}$$

In particular, if $O_{ij} \sim \exp(\lambda_{ij})$ then $A_{ij}(s_1, \ldots, s_n) = \frac{\lambda_{ij}}{\lambda_{ij}+\eta_{ij}(s_1,\ldots,s_n)}$.

The LST $\chi_k(s)$ is obtained by averaging over all possible gate intervals, and by making the following observation. Considering an O_{ij} interval at an arbitrary epoch during that interval, the LST of the joint amounts which have arrived at the queues during the past part of O_{ij} equals

$$\begin{aligned} &\int_0^\infty \mathrm{e}^{-t\eta_{ij}(s_1,\ldots,s_n)} \frac{\mathbb{P}(O_{ij} > t)}{\mathbb{E}[O_{ij}]} \mathrm{d}t \\ &= \frac{1 - A_{ij}(s_1, \ldots, s_n)}{\mathbb{E}[O_{ij}]\eta_{ij}(s_1, \ldots, s_n)}. \end{aligned} \tag{20}$$

This leads to the following result:

$$\begin{aligned} \chi_k(s) &= \frac{\sum_i \sum_j \pi_i p_{ij} \mathbb{E}[O_{ij}] \frac{1-A_{ij}(s,\ldots,s,0,\ldots,0)}{\mathbb{E}[O_{ij}]\eta_{ij}(s,\ldots,s,0,\ldots,0)} \xi_{ki}(s)}{\sum_i \sum_j \pi_i p_{ij} \mathbb{E}[O_{ij}]} \\ &= \frac{\sum_i \sum_j \pi_i p_{ij} \frac{1-A_{ij}(s,\ldots,s,0,\ldots,0)}{\eta_{ij}(s,\ldots,s,0,\ldots,0)} \xi_{ki}(s)}{\sum_i \sum_j \pi_i p_{ij} \mathbb{E}[O_{ij}]}, \end{aligned} \tag{21}$$

where the last s in the n-dimensional expressions in the above formula occurs at position k, and where $\xi_{ki}(s)$, $k \neq i$, are given in (9) and $\xi_{kk}(s)$ in (12). When $O_{ij} \sim \exp(\lambda_{ij})$, (21) becomes

$$\chi_k(s) = \frac{\sum_i \sum_j \frac{\pi_i p_{ij}}{\lambda_{ij}+\eta_{ij}(s,\ldots,s,0,\ldots,0)} \xi_{ki}(s)}{\sum_i \sum_j \frac{\pi_i p_{ij}}{\lambda_{ij}}}. \tag{22}$$

2.4 Multi-dimensional workload distributions

In this subsection we outline how the *joint* workload distribution just after gate openings can be obtained. We provide a fairly detailed procedure for the cases $n = 2$ and $n = 3$ and, for the sake of brevity, under the simplifying assumptions that $p_{ij} = q_j$ for all relevant i and that and that $A_{ij}(\cdot) = A(\cdot)$ for all relevant i, j. For higher values of n, as well as without these simplifying assumptions a similar procedure can be followed; however, it leads to quite messy expressions.

The case $n = 2$
We shall determine the LST of the steady-state joint distribution of the workloads right after gate openings, $\xi(s_1, s_2) = \mathbb{E}[\mathrm{e}^{-s_1 V_1 - s_2 V_2}]$.

If $V_i^{(r)}$ denotes the amount of work in Q_i immediately after the rth gate opening, and $A_i^{(r+1)}$ the amount of fluid entering Q_i between the rth and $(r+1)$st gate openings, then

$$V_1^{(r+1)} = 0, \quad V_2^{(r+1)} = V_1^{(r)} + A_1^{(r+1)} + V_2^{(r)} + A_2^{(r+1)},$$

if the $(r+1)$st gate opening is of gate 1, and

$$V_1^{(r+1} = V_1^{(r)} + A_1^{(r+1)}, \quad V_2^{(r+1)} = 0,$$

if the $(r+1)$st gate opening is of gate 2. In steady state this yields:

$$\xi(s_1, s_2) = q_1 A(s_2, s_2)\xi(s_2, s_2) + q_2 A(s_1, 0)\xi(s_1, 0). \tag{23}$$

Now observe that $\xi(s_1, 0) = \xi_1(s_1)$, and that this term, which only refers to Q_1, can be obtained from the results of Subsection 2.2. Furthermore observe that $\xi(s_2, s_2) = \mathbb{E}[\mathrm{e}^{-s_2 V^{(2)}}]$, a result for the total workload in $Q_1 + Q_2$, which also follows from Subsection 2.2. We are thus able to obtain $\xi(s_1, s_2)$.

Remark 1 Let us assume that station 2 is replaced by L parallel stations $Q_{21}, \ldots, Q_{2L}$. A proportion p_j of every drop that leaves station Q_1 is routed to station Q_{2j} and the gates at station 2 open at the same times. Let $1 - \sum_{j=1}^{L} p_j$ be the proportion that leave the system entirely (from station Q_1). If we denote $V_1, V_{21}, \ldots, V_{2L}$ the workloads in all stations then it is easilly seen that $V_{2j} = p_j V_2$ where (V_1, V_2) was defined in the beginning of this subsection. This immediately implies that the steady-state LST for this case becomes $\xi\left(\alpha_1, \sum_{j=1}^{L} p_j \alpha_{2j}\right)$. This remains true regardless of the simplifying assumptions or the assumptions on the arrival process and in fact, in the n station case, every station can be replaced by parallel stations in a similar manner with the same consequence.

The case $n = 3$
Here we compute the three-dimensional steady-state transform $\xi(s_1, s_2, s_3) = \mathbb{E}[\mathrm{e}^{-s_1V_1 - s_2V_2 - s_3V_3}]$ of workload right after gate openings (under the same simplifying assumptions described in the beginning of this subsection). We have

$$V_1^{(r+1)} = 0, \quad V_2^{(r+1)} = V_1^{(r)} + A_1^{(r+1)} + V_2^{(r)} + A_2^{(r+1)}, \quad V_3^{(r+1)} = V_3^{(r)} + A_3^{(r+1)},$$

if the $(r+1)$st gate opening is of gate 1, and

$$V_1^{(r+1} = V_1^{(r)} + A_1^{(r+1)}, \quad V_2^{(r+1)} = 0, \quad V_3^{(r+1)} = V_2^{(r)} + A_2^{(r+1)} + V_3^{(r)} + A_3^{(r+1)},$$

if the $(r+1)$st gate opening is of gate 2, and

$$V_1^{(r+1} = V_1^{(r)} + A_1^{(r+1)}, \quad V_2^{(r+1)} = V_2^{(r)} + A_2^{(r+1)}, \quad V_3^{(r+1)} = 0,$$

if the $(r+1)$st gate opening is of gate 3. In steady state, this yields:

$$\begin{aligned}\xi(s_1, s_2, s_3) = {} & q_1 A(s_2, s_2, s_3)\xi(s_2, s_2, s_3) + q_2 A(s_1, s_3, s_3)\xi(s_1, s_3, s_3) \\ & + q_3 A(s_1, s_2, 0)\xi(s_1, s_2, 0).\end{aligned} \tag{24}$$

Taking $s_3 = 0$ gives

$$\xi(s_1, s_2, 0) = \frac{q_1 A(s_2, s_2, 0)\xi(s_2, s_2, 0) + q_2 A(s_1, 0, 0)\xi(s_1, 0, 0)}{1 - q_3 A(s_1, s_2, 0)}. \tag{25}$$

Notice that $\xi(s_1, 0, 0) = \mathbb{E}[\mathrm{e}^{-s_1V^{(1)}}]$ and that $\xi(s_2, s_2, 0) = \mathbb{E}[\mathrm{e}^{-s_2V^{(2)}}]$ are known from the previous section, so that $\xi(s_1, s_2, 0)$ is known. Of course, this term is also closely related to the result in (23) for a model with $n = 2$ queues. In fact, a straightforward extension of (23) for the first two queues of an n-queue tandem ASIP is:

$$\begin{aligned}\xi(s_1, s_2, 0, \ldots, 0)) = {} & q_1 A(s_2, s_2, 0, \ldots, 0)\xi(s_2, s_2, 0, \ldots, 0) \\ & + q_2 A(s_1, 0, 0, \ldots, 0)\xi(s_1, 0, 0, \ldots, 0) \\ & + \sum_{j=3}^{n} q_j A(s_1, s_2, 0, \ldots, 0)\xi(s_1, s_2, 0, \ldots, 0)\,.\end{aligned} \tag{26}$$

We still need to determine $\xi(s_2, s_2, s_3)$ and $\xi(s_1, s_3, s_3)$ in (24). Take $s_2 = s_3$ in (24) to get

$$\begin{aligned}\xi(s_1, s_3, s_3) = {} & q_1 A(s_3, s_3, s_3)\xi(s_3, s_3, s_3) + q_2 A(s_1, s_3, s_3)\xi(s_1, s_3, s_3) \\ & + q_3 A(s_1, s_3, 0)\xi(s_1, s_3, 0)\,.\end{aligned} \tag{27}$$

This equation allows us to express $\xi(s_1, s_3, s_3)$ in terms of the, by now known, functions $\xi(s_3, s_3, s_3) = \mathbb{E}[\mathrm{e}^{-s_3V^{(3)}}]$ and $\xi(s_1, s_3, 0)$ (cf. (25)). It remains to determine $\xi(s_2, s_2, s_3)$. For this purpose, take $s_1 = s_2$ in (24):

$$\begin{aligned}\xi(s_2, s_2, s_3) = {} & q_1 A(s_2, s_2, s_3)\xi(s_2, s_2, s_3) + q_2 A(s_2, s_3, s_3)\xi(s_2, s_3, s_3) \\ & + q_3 A(s_2, s_2, 0)\xi(s_2, s_2, 0)\,.\end{aligned} \tag{28}$$

This equation allows us to express $\xi(s_2, s_2, s_3)$ in terms of the, by now known, functions $\xi(s_2, s_3, s_3)$ and $\xi(s_2, s_2, 0) = \mathbb{E}[\mathrm{e}^{-s_2V^{(2)}}]$. Thus, $\xi(s_1, s_2, s_3)$ has been obtained.

3 Model 2: An ASIP with leakage

In this section we consider an ASIP consisting of two stations Q_1 and Q_2 in series, each with its own gate, with the additional feature that there is leakage from Q_1. Namely, the content of Q_1 is not only transferred to Q_2 at openings of the gate after Q_1, but the content also leaks at a fixed rate out of Q_1 (whenever the queue is not empty). We restrict ourselves in this section to gate openings at i.i.d. exponentially distributed intervals, and we assume that the external input processes to the two stations are two independent subordinators (nondecreasing Lévy processes). In Subsection 3.1 we present some preliminary results on a Lévy process reflected at zero, which are used in Subsection 3.2 to derive the steady-state joint workload distribution right after gate openings, right before gate openings and at arbitrary epochs.

3.1 Preliminaries

Let $X = \{X(t)|t \geq 0\}$ be a Lévy process with no negative jumps which is not a subordinator and with Laplace exponent $\varphi(\alpha) = \log \mathbb{E}[\mathrm{e}^{-\alpha X(1)}]$. Denote

$$L_x(t) = -\inf_{0\leq s\leq t}(x + X(s))^- = (L_0(t) - x)^+, \tag{29}$$

$$Z_x(t) = x + X(t) + L_x(t) = X(t) + x \vee L(t)\ , \tag{30}$$

and finally, for $u \geq 0$ let

$$\psi(u) = \inf\{\alpha|\varphi(\alpha) > u\}. \tag{31}$$

Assume that $T \sim \exp(\lambda)$ is independent of X, then for any $\alpha \geq 0$, and $\beta > -\psi(\lambda)$ we have that

$$\mathbb{E}[\mathrm{e}^{-\alpha Z_x(T)-\beta L_x(T)}] = \frac{e^{-\alpha x}(\psi(\lambda)+\beta) - \mathrm{e}^{-\psi(\lambda)x}(\alpha+\beta)}{\left(1-\frac{\varphi(\alpha)}{\lambda}\right)(\psi(\lambda)+\beta)}\ , \tag{32}$$

where for $\alpha = \psi(\lambda)$ the right hand side is defined by continuity via L'Hôpital's rule. This is, in essence, Theorem 3.10 on page 259 of [1], which is a direct application of [12]. It is easy to check that the proof is valid for all α, β as given here and not just $\alpha, \beta > 0$ as in [1]. This will be important later.

Whenever $Y = \{Y(t)|t \geq 0\}$ is a measurable process which is independent of $T_\lambda \sim \exp(\lambda)$, then clearly, for every $\gamma > -\lambda$ we have that

$$\mathbb{E}[Y(T_\lambda)e^{-\gamma T_\lambda}] = \int_0^\infty \mathbb{E}[Y(t)]\mathrm{e}^{-\gamma t}\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t = \frac{\lambda}{\lambda+\gamma}\mathbb{E}[Y(T_{\lambda+\gamma})]\ . \tag{33}$$

From (32) and (33) it immediately follows that for each $\alpha \geq 0$, $\gamma > -\lambda$ and $\beta > -\psi(\lambda+\gamma)$:

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-\alpha Z_x(T)-\beta L_x(T)-\gamma T}] &= \frac{\lambda}{\lambda+\gamma}\cdot\frac{\mathrm{e}^{-\alpha x}(\psi(\lambda+\gamma)+\beta) - \mathrm{e}^{-\psi(\lambda+\gamma)x}(\alpha+\beta)}{\left(1-\frac{\varphi(\alpha)}{\lambda+\gamma}\right)(\psi(\lambda+\gamma)+\beta)} \\ &= \frac{\mathrm{e}^{-\alpha x}(\psi(\lambda+\gamma)+\beta) - \mathrm{e}^{-\psi(\lambda+\gamma)x}(\alpha+\beta)}{\left(1+\frac{\gamma-\varphi(\alpha)}{\lambda}\right)(\psi(\lambda+\gamma)+\beta)}\ .\end{aligned} \tag{34}$$

3.2 Analysis

Now consider a system consisting of two stations in series. The external input process of station Q_i is a nondecreasing Lévy process J_i, $i = 1, 2$. These are independent subordinators with

$$\eta_i(\alpha) = -\log \mathrm{e}^{-\alpha J_i(1)} = c_i\alpha + \int_{(0,\infty)} (1 - \mathrm{e}^{-\alpha u})\nu_i(\mathrm{d}u),$$

where $c_i \geq 0$ and ν_i is a Lévy measure satisfying $\int_{(0,\infty)}(u \wedge 1)\nu_i(\mathrm{d}u) < \infty$ and $\nu_i(-\infty, 0] = 0$.

The cumulative input to Q_1 is $x_1 + J_1(t)$ where $x_1 \geq 0$ is its initial state. Whenever Q_1 is not empty, the content leaks (is processed) at some rate $r \geq 0$. When $c_1 < r$ and Q_1 is empty, the leak is at rate c_1. A proportion $p \in [0, 1]$ leaks into Q_2 and the rest leaves the system altogether. At independent intervals, also independent of J_1, J_2 and distributed $\exp(\lambda_1)$, the entire content of Q_1 is transferred to Q_2. As for Q_2, the cumulative input is whatever arrives from Q_1 (either from the leakage or from the occasional transfer) as well as $x_2 + J_2(t)$ where $x_2 \geq 0$ is the initial state of Q_2. At independent intervals which are distributed $\exp(\lambda_2)$ and independent of everything else (including the inter-transfer times of Q_1) all the available content of Q_2 leaves the system all at once. This is the two-queue ASIP system that we would like to explore.

Denote $\lambda = \lambda_1 + \lambda_2$. For $X_1(t) = J_1(t) - rt$, with Laplace-Stieltjes exponent $\varphi_1(\alpha) = \log\mathbb{E}[\mathrm{e}^{-\alpha X_1(1)}] = r\alpha - \eta_1(\alpha)$, the content of Q_1 at time $t \geq 0$ is $Z_{1,x_1}(t)$ where Z_{1,x_1} replaces Z_x in (30). As long as there is no transfer until time $t \geq 0$, the input to Q_2 is $p(rt - L_{1,x_1}(t))$, as $L_{1,x_1}(t)$ is the cumulative lost capacity.

If $T \sim \exp(\lambda)$ (the minimum of the transfer times from Q_1 and Q_2), then with probability λ_1/λ there is a transfer from Q_1 to Q_2, in which case the state of the stations will be $(0, Z_{1,x_1}(T) + p(rT - L_{1,x_1}(T)) + x_2 + J_2(T))$, and with probability λ_2/λ the state will be $(Z_{1,x_1}(T), 0)$. Therefore, we will be interested in the LST of $Z_{1,x_1}(T)+p(rT-L_{1,x_1}(T))+J_2(T)$ and that of $Z_{1,x_1}(T)$.

If $r \leq c_1$ (which includes the case $r = 0$) then X_1 is a subordinator and (34) does not apply. However, in this case the result is far simpler, as $L_x(T) = 0$ and $Z_{1,x_1}(T) = x_1+X_1(T)$. When $r > c_1$ then, noting that

$$\mathbb{E}[\mathrm{e}^{-\alpha(Z_{1,x_1}(T)+p(rT-L_{1,x_1}(T))+J_2(T))}] = \mathbb{E}[\mathrm{e}^{-(\alpha Z_{1,x_1}(T)-\alpha p L_{1,x_1}(T)+(pr\alpha+\eta_2(\alpha))T)}]\ , \tag{35}$$

we simply apply (34) setting either $\gamma = pr\alpha + \eta_2(\alpha)$ and $\beta = -p\alpha$ (for $Z_{1,x_1}(T) + p(rT - L_{1,x_1}(T)) + J_2(T)$) or $\gamma = \beta = 0$ (for $Z_{1,x_1}(T)$). Recall that in order to use (34) we must have $\beta > -\psi_1(\lambda + \gamma)$, cf. (31). To see that this holds in this case, note that, when $\beta < 0$, this is equivalent to $\varphi_1(-\beta) < \lambda + \gamma$ and if we insert $\beta = -p\alpha$ and $\gamma = pr\alpha + \eta_2(\alpha)$ and observe that $\varphi_1(\alpha) = r\alpha - \eta_1(\alpha)$, then indeed, as required,

$$\varphi_1(-\beta) = rp\alpha - \eta_1(p\alpha) < \lambda + pr\alpha + \eta_2(\alpha) = \lambda + \gamma. \tag{36}$$

This system is regenerative. The reason is that starting from (x_1, x_2) the system will (almost surely) reach some state $(0, x_2')$ after which it will reach a state $(Z_{1,0}(\tau), 0)$ where τ is the first time thereafter that the second station is emptied. This state is a regenerative one and it is clear that the inter-regeneration times distribution has a finite mean and is nonarithmetic; it actually has a density. Thus, there exists a limiting=ergodic=stationary distribution for the joint content process. Assume that (Z_1, Z_2) has this joint distribution (of the buffer contents

right *after* an arbitrary gate opening) and denote $f_A(\alpha_1,\alpha_2)=\mathbb{E}[\mathrm{e}^{-\alpha_1 Z_1-\alpha_2 Z_2}]$. Then given the preceding arguments, we must have for the case where $r>c_1$, that

$$f_A(\alpha_1,\alpha_2) \tag{37}$$
$$= \frac{\lambda_1}{\lambda}\frac{f_A(\alpha_2,\alpha_2)(\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2))-p\alpha_2)-f_A(\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2)),\alpha_2)\alpha_2(1-p)}{\left(1-\frac{(1-p)r\alpha_2-\eta_1(\alpha_2)-\eta_2(\alpha_2)}{\lambda}\right)(\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2))-p\alpha_2)}$$
$$+\frac{\lambda_2}{\lambda}\frac{f_A(\alpha_1,0)\psi_1(\lambda)-f_A(\psi_1(\lambda),0)\alpha_1}{\left(1-\frac{r\alpha_1-\eta_1(\alpha_1)}{\lambda}\right)\psi_1(\lambda)}\ . \tag{38}$$

We shall successively determine (i) $f_A(\alpha_1,0)$ and $f_A(\psi_1(\lambda),0)$, (ii) $f_A(\alpha_2,\alpha_2)$, $f_A(\psi_1(\lambda+pr\alpha_2),\alpha_2)$ and $f_A(\alpha_1,\alpha_2)$.
(i) *Determination of* $f_A(\alpha_1,0)$ *and* $f_A(\psi_1(\lambda),0)$.
Taking $\alpha_2=0$ in (38), with $\alpha_1\neq\psi_1(\lambda_1)$, gives:

$$f_A(\alpha_1,0)=\frac{\lambda_1}{\lambda}+\frac{\lambda_2}{\lambda}\frac{f_A(\alpha_1,0)\psi_1(\lambda)-f_A(\psi_1(\lambda),0)\alpha_1}{1-\frac{r\alpha_1-\eta_1(\alpha_1)}{\lambda}\psi_1(\lambda)}, \tag{39}$$

which is equivalent to

$$(\lambda_1-r\alpha_1+\eta_1(\alpha_1))f_A(\alpha_1,0)=\lambda_1-\lambda f_A(\psi_1(\lambda),0)\frac{\alpha_1}{\psi_1(\lambda)}. \tag{40}$$

Setting $\alpha_1=\psi_1(\lambda_1)$, the lefthand side of (40) becomes zero, and hence also the righthand side should be zero, implying $f_A(\psi_1(\lambda),0)=\frac{\lambda_1}{\lambda}\frac{\psi_1(\lambda)}{\psi_1(\lambda_1)}$, and hence

$$f_A(\alpha_1,0)=\frac{\lambda_1(1-\frac{\alpha_1}{\psi_1(\lambda_1)})}{\lambda_1-r\alpha_1+\eta_1(\alpha_1)}=\frac{1-\frac{\alpha_1}{\psi_1(\lambda_1)}}{1-\frac{\varphi_1(\alpha_1)}{\lambda_1}}. \tag{41}$$

Notice that this is also the LST of the workload in Q_1 just before a gate opening of Q_1, i.e., after an $\exp(\lambda_1)$ amount of time starting from an empty state, cf. Theorem 4.1 of [6]. This is expected by PASTA.

(ii) *Determination of* $f_A(\alpha_2,\alpha_2)$, $f_A(\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2)),\alpha_2)$ *and* $f_A(\alpha_1,\alpha_2)$.
Introducing the following functions for terms appearing in (38):

$$A(\alpha_2)=\frac{\lambda_1}{\lambda}\frac{1}{1-\frac{(1-p)r\alpha_2-\eta_1(\alpha_2)-\eta_2(\alpha_2)}{\lambda}}, \tag{42}$$

$$B(\alpha_2)=-A(\alpha_2)\frac{\alpha_2(1-p)}{\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2))-p\alpha_2}, \tag{43}$$

$$H(y)=\frac{\lambda_2}{\lambda}\frac{f(y,0)\psi_1(\lambda)-f(\psi_1(\lambda),0)y}{(1-\frac{ry-\eta_1(y)}{\lambda})\psi_1(\lambda)}, \tag{44}$$

one can rewrite (38) as

$$f_A(\alpha_1,\alpha_2)=H(\alpha_1)+G(\alpha_2), \tag{45}$$

where

$$G(\alpha_2)=A(\alpha_2)f_A(\alpha_2,\alpha_2)+B(\alpha_2)f_A(\psi_1(\lambda+pr\alpha_2+\eta_2(\alpha_2)),\alpha_2). \tag{46}$$

The decomposition in (45) makes sense as we are observing the system just after gate openings. With probability λ_i/λ, the gate opening was at Q_i, and then Q_i has become empty, yielding a term without α_i, $i = 1, 2$. The decomposition form of course implies that $\mathbb{E}[Z_1 Z_2] = 0$, which obviously holds because after each gate opening at least one of the two queues has become empty, hence $Z_1 Z_2 = 0$. This also implies, as is quite intuitive, that Z_1 and Z_2 are negatively correlated.

$G(\alpha_2)$ is determined by substituting $\alpha_1 = \alpha_2$ respectively $\alpha_1 = \psi_1(\lambda + pr\alpha_2 + \eta_2(\alpha_2))$ in (45):

$$G(\alpha_2) = \frac{A(\alpha_2)H(\alpha_2) + B(\alpha_2)H(\psi_1(\lambda + pr\alpha_2 + \eta_2(\alpha_2)))}{1 - A(\alpha_2) - B(\alpha_2)},$$

and hence

$$f_A(\alpha_1, \alpha_2) = H(\alpha_1) + \frac{A(\alpha_2)H(\alpha_2) + B(\alpha_2)H(\psi_1(\lambda + pr\alpha_2 + \eta_2(\alpha_2)))}{1 - A(\alpha_2) - B(\alpha_2)}. \tag{47}$$

For completeness we give the LST of the total workload in the two queues, $f_A(\alpha, \alpha)$:

$$f_A(\alpha, \alpha) = \frac{(1 - B(\alpha))H(\alpha) + B(\alpha)H(\psi_1(\lambda + pr\alpha + \eta_2(\alpha)))}{1 - A(\alpha) - B(\alpha)}. \tag{48}$$

The time-stationary workload LST

Above we have computed the steady-state joint workload LST of our system just after gate openings. If $f_B(\alpha_1, \alpha_2)$ is the steady-state joint workload LST just *before* (any) gate openings, then by PASTA it is also the continuous-time steady-state workload LST. Clearly, one relationship between f_A and f_B is as follows:

$$f_A(\alpha_1, \alpha_2) = \frac{\lambda_1}{\lambda} f_B(\alpha_2, \alpha_2) + \frac{\lambda_2}{\lambda_1} f_B(\alpha_1, 0)\,. \tag{49}$$

However, this is not enough and in order to compute f_B we need to compute the joint LST of the system that starts with distribution having LST f_A and ends after an independent exponential time period with parameter λ. Thus, letting $T \sim \exp(\lambda)$ be independent of everything else, then in an identical manner as for (37) we have (when $r > c_1$) that

$$\begin{aligned} &f_B(\alpha_1, \alpha_2) \qquad (50)\\ &= \frac{f_A(\alpha_1, \alpha_2)(\psi_1(\lambda + pr\alpha_2 + \eta_2(\alpha_2)) - p\alpha_2) - f_A(\psi_1(\lambda + pr\alpha_2 + \eta_2(\alpha_2)), \alpha_2)(\alpha_1 - p\alpha_2)}{\left(1 + \frac{rp\alpha_2 + \eta_2(\alpha_2) - (r\alpha_1 - \eta_1(\alpha_1))}{\lambda}\right)(\psi_1(\lambda + rp\alpha_2 + \eta_2(\alpha_2)) - p\alpha_2)}\,. \end{aligned}$$

Therefore, we also have f_B, the steady-state joint workload LST just before gate openings and at arbitrary epochs.

Determination of moments

It readily follows from (41) that the mean buffer content in Q_1 right after an arbitrary gate opening is given by

$$\mathbb{E}[Z_1] = -\frac{d}{d\alpha_1} f_A(\alpha_1, 0)|_{\alpha_1 = 0} = \frac{\eta_1'(0) - r}{\lambda_1} + \frac{1}{\psi_1(\lambda_1)}. \tag{51}$$

$\mathbb{E}[Z_2]$ follows by differentiating the expression in (45) w.r.t. α_2. Alternatively, we could obtain $\mathbb{E}[Z_1 + Z_2]$ from (48) and then subtract $\mathbb{E}[Z_1]$. Omitting the messy details, the end result can be written as follows:

$$\begin{aligned}\mathbb{E}[Z_2] &= \frac{\eta_1'(0) - (1-p)r}{\lambda_2} + \frac{\lambda_1}{\lambda\lambda_2}\eta_2'(0) + \frac{\lambda_1^2}{\lambda\lambda_2}\frac{1-p}{\psi_1(\lambda)} \\ &+ \frac{\lambda_1}{\lambda}(\mathbb{E}[Z_1] + \frac{\lambda_1}{\lambda_2}(\frac{\psi_1(\lambda)}{\psi_1(\lambda_1)} - 1)\frac{1}{\psi_1(\lambda)} + \frac{\eta_1'(0) - r}{\lambda_1}) \\ &+ \frac{\lambda_1}{\lambda_2}\frac{1-p}{\psi_1(\lambda)}H(\psi_1(\lambda)). \end{aligned} \tag{52}$$

4 Model 3: The shot-noise counterpart of Model 2

In this section we again consider an ASIP model consisting of two queues Q_1 and Q_2 in series, Q_i having a gate which opens at independent, $\exp(\lambda_i)$ distributed intervals, for $i = 1, 2$. If the gate of Q_1 opens, the buffer content of this queue instantaneously moves to Q_2; if the gate of Q_2 opens, the buffer content of this queue instantaneously leaves the system. Again, the two queues receive external input according to two independent Lévy subordinators J_i, $i = 1, 2$. As in Model 2, there is leakage from Q_1; a fraction p of the leakage from Q_1 moves to Q_2 and the rest disappears altogether. So far the model description is the same as for Model 2 in Section 3. The special feature of the present model, compared to Model 2, is that in between gate openings, the workload at Q_1 behaves like a *shot-noise process*. In a shot-noise process, the workload decreases proportional to the buffer content, at rate rx when the buffer content equals x; this amounts to an exponentially decreasing process, and can be seen as a fluid-type counterpart of an infinite-server queue. It can model situations in which all material that is present in a station is processed simultaneously.

Material from Q_2 can leave this queue only when it has a gate opening. For this case, we note that when there is no gate opening before time t, then the two buffer contents $Z_{1,x_1}(t)$ and $Z_{2,x_1,x_2}(t)$ evolve as follows:

$$\begin{aligned} Z_{1,x_1}(t) &= x_1 + J_1(t) - r\int_0^t Z_{1,x_1}(s)ds = x_1 e^{-rt} + \int_{(0,t]} e^{-r(t-s)}\mathrm{d}J_1(s), \\ Z_{2,x_1,x_2}(t) &= p(x_1 + J_1(t) - Z_{1,x_1}(t)) + x_2 + J_2(t) \\ &= p\left(x_1(1-e^{-rt}) + \int_{(0,t]}(1-e^{-r(t-s)})\mathrm{d}J_1(s)\right) + x_2 + J_2(t)\,. \end{aligned} \tag{53}$$

Indeed, at any time t (before the first gate opening) a fraction p of the difference between $x_1 + J_1(t)$ and the buffer content $Z_{1,x_1}(t)$ has moved to Q_2. Recalling that for nonnegative Borel functions h we have (see, e.g., Formula (8) of [4]),

$$\mathbb{E}[\mathrm{e}^{-\int_{(0,t]} h(t-s)\mathrm{d}J_1(s)}] = \mathrm{e}^{-\int_0^t \eta_1(h(s))\mathrm{d}s}, \tag{54}$$

this implies that

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-\alpha_1 Z_{1,x_1}(t)-\alpha_2 Z_{2,x_1,x_2}(t)}] &= \exp\Bigg(-\alpha_2 x_2 - x_1(\alpha_1 e^{-rt} + \alpha_2 p(1-e^{-rt})) \\ &\quad - \int_0^t \eta_1(\alpha_1 \mathrm{e}^{-rs} + \alpha_2 p(1-e^{-rs}))\mathrm{d}s - \eta_2(\alpha_2)t\Bigg) \qquad (55)\\ &= \exp\Bigg(-\alpha_2 x_2 - x_1(\alpha_1 \mathrm{e}^{-rt} + \alpha_2 p(1-e^{-rt})) \\ &\quad - \int_{\mathrm{e}^{-rt}}^1 \eta_1(\alpha_1 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{ru} - \eta_2(\alpha_2)t\Bigg).\end{aligned}$$

Multiplying by $\lambda \mathrm{e}^{-\lambda t}$, where $\lambda = \lambda_1 + \lambda_2$, and integrating gives, after the obvious change of variables $v = \mathrm{e}^{-rt}$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{-\alpha_1 Z_{1,x_1}(T)-\alpha_2 Z_{2,x_1,x_2}(T)}] &= \frac{\lambda}{r}\int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1}\exp\Bigg(-\alpha_2 x_2 - x_1(\alpha_1 v + \alpha_2 p(1-v)) \\ &\quad - \int_v^1 \eta_1(\alpha_1 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{ru}\Bigg)\mathrm{d}v\ , \qquad (56)\end{aligned}$$

where $T \sim \exp(\lambda)$ is independent of everything else.

Remark 2 If in addition we assume that J_1 is a compound Poisson process with arrival rate λ and jumps $\sim \exp(\mu)$, then

$$\eta_1(\alpha) = \lambda\left(1 - \frac{\mu}{\mu+\alpha}\right) = \frac{\lambda\alpha}{\mu+\alpha}, \qquad (57)$$

in which case

$$\int_v^1 \eta_1(\alpha_1 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{ru} \qquad (58)$$

can be computed explicitly by observing that

$$\frac{\eta_1(\alpha_1 u + \alpha_2 p(1-u))}{ru} = \frac{\lambda}{r(\mu+p\alpha_2)}\left(\frac{p\alpha_2}{u} + \frac{\mu(\alpha_1 - p\alpha_2)}{(\alpha_1-p\alpha_2)u + \mu + p\alpha_2}\right), \qquad (59)$$

and so the integral in (58) becomes

$$\frac{\lambda}{r(\mu+p\alpha_2)}\left(-p\alpha_2 \log v + \mu \log\left(\frac{\alpha_1+\mu}{\alpha_1 v + p\alpha_2(1-v)+\mu}\right)\right). \qquad (60)$$

Multiplying by minus one and taking the exponent gives

$$v^{\frac{\lambda p\alpha_2}{r(\mu+p\alpha_2)}}\left(\frac{\alpha_1+\mu}{\alpha_1 v + p\alpha_2(1-v)+\mu}\right)^{-\frac{\lambda\mu}{r(\mu+p\alpha_2)}}. \qquad (61)$$

We shall now determine the steady-state joint workload LST just before gate openings and just after gate openings. Let $(Z_{1,A}, Z_{2,A})$ denote the steady-state workload vector at Q_1 and Q_2 just *after* an arbitrary gate opening, with LST $F_A(\alpha_1, \alpha_2)$, and let $(Z_{1,B}, Z_{2,B})$ denote the steady-state workload vector just *before* an arbitrary gate opening, with LST $F_B(\alpha_1, \alpha_2)$. Observe that, if gate 1 just opened, then $Z_{1,A}$ becomes zero and $Z_{2,A}$ becomes $Z_{1,B} + Z_{2,B}$; and if gate 2 just opened, then $Z_{1,A}$ becomes $Z_{1,B}$ and $Z_{2,A}$ becomes 0. Hence, with (again) $\lambda = \lambda_1 + \lambda_2$:

$$F_A(\alpha_1, \alpha_2) = \frac{\lambda_1}{\lambda} F_B(\alpha_2, \alpha_2) + \frac{\lambda_2}{\lambda} F_B(\alpha_1, 0). \tag{62}$$

From (56), conditioning on $Z_{1,A} = x_1, Z_{2,A} = x_2$, we have

$$F_B(\alpha_1, \alpha_2) = \frac{\lambda}{r} \int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1} \mathrm{e}^{-\frac{1}{r}\int_v^1 \eta_1(\alpha_1 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{u}} F_A(\alpha_1 v + \alpha_2 p(1-v), \alpha_2) \mathrm{d}v. \tag{63}$$

We shall first determine $K_1(\alpha_1) = F_B(\alpha_1, 0)$. Taking $\alpha_2 = 0$ in (63) yields

$$\begin{aligned} K_1(\alpha_1) &= \frac{\lambda}{r} \int_0^1 v^{\frac{\lambda}{r}-1} \mathrm{e}^{-\frac{1}{r}\int_v^1 \eta_1(\alpha_1 u)\frac{\mathrm{d}u}{u}} F_A(\alpha_1 v, 0) \mathrm{d}v \\ &= \frac{\lambda}{r} \int_0^{\alpha_1} \frac{y^{\frac{\lambda}{r}-1}}{\alpha_1^{\frac{\lambda}{r}}} \mathrm{e}^{-\frac{1}{r}\int_y^{\alpha_1} \eta_1(z)\frac{\mathrm{d}z}{z}} \left(\frac{\lambda_1}{\lambda} + \frac{\lambda_2}{\lambda} K_1(y) \right) \mathrm{d}y. \end{aligned} \tag{64}$$

Differentiation w.r.t. α_1 results in a first-order inhomogeneous differential equation:

$$K_1'(\alpha_1) = \frac{\lambda_1}{r\alpha_1} - \left(\frac{\lambda_1}{r\alpha_1} + \frac{\eta_1(\alpha_1)}{r\alpha_1} \right) K_1(\alpha_1), \tag{65}$$

whose solution is readily seen to be

$$K_1(\alpha_1) = \alpha_1^{-\frac{\lambda_1}{r}} \mathrm{e}^{-\frac{1}{r}\int_0^{\alpha_1} \frac{\eta_1(u)}{u}\mathrm{d}u} \left[C + \frac{\lambda_1}{r} \int_0^{\alpha_1} v^{\frac{\lambda_1}{r}-1} \mathrm{e}^{\frac{1}{r}\int_0^v \frac{\eta_1(u)}{u}\mathrm{d}u} \mathrm{d}v \right]. \tag{66}$$

The fact that we should have $K_1(0) = 1$ implies that the term between square brackets should be zero, and hence $C = 0$. One can subsequently quickly verify, by the transformation $w = v/\alpha_1$, that

$$\lim_{\alpha_1 \downarrow 0} K_1(\alpha_1) = \frac{\lambda_1}{r} \int_0^1 w^{\frac{\lambda_1}{r}-1} \mathrm{d}w = 1. \tag{67}$$

We conclude that

$$K_1(\alpha_1) = F_B(\alpha_1, 0) = \frac{\lambda_1}{r\alpha_1} \int_0^{\alpha_1} \left(\frac{v}{\alpha_1} \right)^{\frac{\lambda_1}{r}-1} \mathrm{e}^{-\frac{1}{r}\int_v^{\alpha_1} \frac{\eta_1(u)}{u}\mathrm{d}u} \mathrm{d}v. \tag{68}$$

By PASTA, this is also the LST of the steady-state workload in Q_1 at an arbitrary epoch. We next turn to the determination of $K_2(\alpha_2) = F_B(\alpha_2, \alpha_2)$, which, again by PASTA, is the LST of the steady-state total workload in the ASIP system at an arbitrary epoch. Taking $\alpha_1 = \alpha_2$ in (63) gives

$$\begin{aligned} K_2(\alpha_2) &= \frac{\lambda}{r} \int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1} \mathrm{e}^{-\frac{1}{r}\int_v^1 \eta_1(\alpha_2 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{u}} F_A(\alpha_2 v + \alpha_2 p(1-v), \alpha_2) \mathrm{d}v \\ &= \frac{\lambda}{r} \int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1} \\ &\times \mathrm{e}^{-\frac{1}{r}\int_v^1 \eta_1(\alpha_2 u + \alpha_2 p(1-u))\frac{\mathrm{d}u}{u}} \left(\frac{\lambda_1}{\lambda} K_2(\alpha_2) + \frac{\lambda_2}{\lambda} K_1(\alpha_2 p + \alpha_2(1-p)v) \right) \mathrm{d}v. \end{aligned} \tag{69}$$

Solving for $K_2(\alpha_2)$ yields

$$\begin{aligned} K_2(\alpha_2) &= [1-\frac{\lambda_1}{r}\int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1}\mathrm{e}^{-\frac{1}{r}\int_v^1\eta_1(\alpha_2 u+\alpha_2 p(1-u))\frac{\mathrm{d}u}{u}}]^{-1} \\ &\times \frac{\lambda_2}{r}\int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)}{r}-1}\mathrm{e}^{-\frac{1}{r}\int_v^1\eta_1(\alpha_2 u+\alpha_2 p(1-u))\frac{\mathrm{d}u}{u}}K_1(\alpha_2 p+\alpha_2(1-p)v)\mathrm{d}v. \end{aligned} \tag{70}$$

One could subsequently substitute the expression for $K_1(\alpha_1)$ as found in (66) in (70). This results in a quite complicated integral, which it seems that one has to evaluate numerically. However, if $p=1$ (so all the leakage goes to Q_2) then (70) simplifies:

$$\begin{aligned} K_2(\alpha_2) &= [1-\frac{\lambda_1}{r}\int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)+\eta_1(\alpha_2)}{r}-1}\mathrm{d}v]^{-1}\frac{\lambda_2}{r}\int_0^1 v^{\frac{\lambda+\eta_2(\alpha_2)+\eta_1(\alpha_2)}{r}-1}K_1(\alpha_2)\mathrm{d}v \\ &= \frac{\frac{\lambda_2}{\lambda+\eta_2(\alpha_2)+\eta_1(\alpha_2)}K_1(\alpha_2)}{1-\frac{\lambda_1}{\lambda+\eta_2(\alpha_2)+\eta_1(\alpha_2)}} = \frac{\lambda_2}{\lambda_2+\eta_2(\alpha_2)+\eta_1(\alpha_2)}K_1(\alpha_2). \end{aligned} \tag{71}$$

Remark 3 When the initial workloads at the two stations are x_1, x_2 and when $p=1$, the sum of the workloads at the two stations just before a gate opening is clearly $x_1+x_2+J_1(T)+J_2(T)$. Therefore, it is obvious that we necessarily have that

$$\mathbb{E}[\mathrm{e}^{-\alpha Z_{1,x_1}(T)-\alpha Z_{2,x_1,x_2}(T)}] = \mathrm{e}^{-\alpha x_1-\alpha x_2}\frac{\lambda}{\lambda+\eta_1(\alpha)+\eta_2(\alpha)}\,. \tag{72}$$

This agrees with (56) upon setting $\alpha_1=\alpha_2=\alpha$ (and $p=1$) as well as formula (63), which reduces to the following, when we take $p=1$ and $\alpha_1=\alpha_2=\alpha$:

$$F_B(\alpha,\alpha) = \frac{\lambda}{\lambda+\eta_1(\alpha)+\eta_2(\alpha)}F_A(\alpha,\alpha). \tag{73}$$

In combination with (62), this readily agrees with (71).

References

[1] S. Asmussen (2003). *Applied Probability and Queues: 2nd ed.* Springer, New York.

[2] R. A. Blythe and M. R. Evans (2007). Nonequilibrium steady states of matrix product form: A solver's guide. *J. Phys. A: Math. Theor.* **40**, R333-R441.

[3] O.L. Bonomo and S. Reuveni (2019). Occupancy correlations in the asymmetric simple inclusion process. *Physical Review E.* **100**, 042109.

[4] O.J. Boxma, O. Kella and M.R.H. Mandjes (2019). Infinite-server systems with Coxian arrivals. *Queueing Systems* **92**, 233-255.

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

[6] K. Debicki and M.R.H. Mandjes (2015). *Queues and Lévy Fluctuation Theory.* Springer, New York.

[7] B. Derrida (1998). An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Physics Reports* **301**, 65-83.

[8] O. Golinelli and K. Mallick (2006). The asymmetric simple exclusion process: an integrable model for non-equilibrium statistical mechanics. *J. Phys. A: Math. Gen.* **39**, 12679-12705.

[9] K. Heckmann (1972). Biomembranes 3, edited by F. Kreuzer and J. F. G. Slegers. Plenum, New York.

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

[11] R.R.P. Jackson (1956). Random queueing processes with phase-type service. J. *Royal Statistical Society Series B (Methodological)* **18**, 129-132.

[12] O. Kella and W. Whitt (1992). Useful martingales for stochastic storage processes with Lévy input. *J. Appl. Probab.* **29**, 396-403.

[13] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin (1968). Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* **6**, 1-5.

[14] S. Reuveni (2014). *Tandem Stochastic Systems: The Asymmetric Simple Inclusion Process*. PhD Thesis, Tel-Aviv University.

[15] S. Reuveni, I. Eliazar and U. Yechiali (2011). Asymmetric inclusion process. *Physical Review* **E 84**, 041101, 1-16.

[16] S. Reuveni, I. Eliazar and U. Yechiali (2012). Limit laws for the asymmetric inclusion process. *Physical Review* **E 86**, 061133, 1-17.

[17] S. Reuveni, I. Eliazar and U. Yechiali (2012). Asymmetric inclusion process as a showcase of complexity. *Physical Review Letters* **109**, 020603, 1-4.

[18] S. Reuveni, O. Hirschberg, I. Eliazar and U. Yechiali (2014). Occupation probabilities and fluctuations in the asymmetric inclusion process. *Physical Review* **E 89**, 042109, 1-23.