Asymmetric inclusion process

Shlomi Reuveni,[1,2] Iddo Eliazar,[3] and Uri Yechiali[2]
[1]*School of Chemistry, Tel-Aviv University, Tel-Aviv IL-69978, Israel*
[2]*Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv IL-69978, Israel*
[3]*Department of Technology Management, Holon Institute of Technology, Holon IL-58102, Israel*
(Received 3 June 2011; revised manuscript received 28 August 2011; published 3 October 2011)

We introduce and explore the asymmetric inclusion process (ASIP), an exactly solvable bosonic counterpart of the fermionic asymmetric exclusion process (ASEP). In both processes, random events cause particles to propagate unidirectionally along a one-dimensional lattice of n sites. In the ASEP, particles are subject to *exclusion* interactions, whereas in the ASIP, particles are subject to *inclusion* interactions that coalesce them into inseparable clusters. We study the dynamics of the ASIP, derive evolution equations for the mean and probability generating function (PGF) of the sites' occupancy vector, obtain explicit results for the above mean at steady state, and describe an iterative scheme for the computation of the PGF at steady state. We further obtain explicit results for the load distribution in steady state, with the load being the total number of particles present in all lattice sites. Finally, we address the problem of load optimization, and solve it under various criteria. The ASIP model establishes bridges between statistical physics and queueing theory as it represents a tandem array of queueing systems with (unlimited) batch service, and a tandem array of growth-collapse processes.

DOI: 10.1103/PhysRevE.84.041101 PACS number(s): 02.50.Ey, 05.40.−a

I. INTRODUCTION

In this paper, we introduce and explore an exactly solvable lattice-gas model in one dimension: the *asymmetric inclusion process* (ASIP). Joining a recent gallery of innovative research papers [1–4], which establish bridges between statistical physics and queueing theory, the ASIP model describes (i) a "bosonic" counterpart of the "fermionic" asymmetric exclusion process (ASEP) [5,6], (ii) a tandem array of queueing systems with batch service [7,8], and (iii) a tandem array of growth-collapse processes [9,10].

A. ASEP

The ASEP, a stochastic process taking place on a discrete one-dimensional lattice of n sites, plays a paradigmatic role in nonequilibrium statistical physics. The ASEP has a long history, having first appeared in the literature as a model of biopolymerization [11] and transport across membranes [12]. Over the years, the ASEP and models that resemble it in spirit were used to study a wide range of physical phenomena: transport of macromolecules through thin vessels [13], hopping conductivity in solid electrolytes [14], reptation of polymer in a gel [15], traffic flow [16], gene translation [17,18], surface growth [19,20], sequence alignment [21], molecular motors [22], and the directed motion of tracer particles in the presence of dynamical backgrounds [23–26].

The ASEP serves as a model for a unidirectionally driven lattice gas of particles subject to *exclusion* interactions. Particles are fed, randomly in time, into the leftmost site of a one-dimensional lattice and propagate unidirectionally (to the right) through the lattice. Particles hop from each site to its right-neighboring site randomly in time, with the hopping restricted by the exclusion principle, which allows sites to be occupied by no more than one particle at a time. At the rightmost site, particles exit the system randomly in time. The random inflow into the leftmost site, the random instants of hopping from site to site, and the random outflow from the rightmost site are all governed by independent "exponential clocks" with given rates. The exclusion principle causes jamming throughout the lattice, and renders the ASEP dynamics highly nontrivial. Despite its simple description and its one dimensionality, the ASEP displays a complex and intricate behavior [5,6,27].

B. ASIP

The exclusion principle is central to the ASEP. While this principle is often suitable for the description of the physical scenario at hand, this is not always the case. Altering the ASEP such that arbitrarily many particles are allowed to simultaneously occupy any given site, one ends up with two different models: the *tandem Jackson network* (which we shall discuss momentarily) and the ASIP. The ASIP is similar to the ASEP, albeit replacing the fermionic *exclusion* principle by a bosonic *inclusion* principle. The ASIP's inclusion principle allows each site to be occupied by an arbitrary number of particles at the same time, and all particles that simultaneously occupy a site are "glued" together into inseparable particle clusters that move together to the next site (or out of the system, in the case of the rightmost site). The mathematical details of the ASIP model are further described in Sec. II of this paper.

C. Other bosonic models

Other bosonic models have been studied in the past. A noteworthy representative is the tandem Jackson network. The tandem Jackson network [28–30] is a sequential array of n service stations. Jobs arrive at the leftmost station randomly in time. At each station, (i) arriving jobs queue up in line (according to their order of arrival) and await service; (ii) only one job is served at a time, and the service durations are governed by exponential clocks; (iii) after service, jobs

TABLE I. Occupancy-service classification of one-dimensional lattice-gas models. The columns specify the occupancy capacity per site (the upper bound for the number of particles (jobs) allowed to simultaneously occupy a given site), and the rows specify the service capacity per site (the upper bound for the size of the particle clusters (job batches) served simultaneously by a given site). If the occupancy capacity is unity, then the ASEP model is attained (regardless of the service capacity). An unlimited occupancy capacity coupled with a unit service capacity yields tandem Jackson networks. Unlimited occupancy and service capacities yield the bosonic batch service ASIP model.

	Occupancy capacity: 1	Occupancy capacity: ∞
Service capacity: 1	ASEP	Tandem Jackson
Service capacity: ∞	ASEP	ASIP

move on to the next station (to the right). After service at the rightmost station, jobs leave the system. The translation between the aforementioned queueing theory setting and a physical setting is straightforward: jobs are particles and service stations are sites. From a queueing perspective, the ASIP is a tandem Jackson network with *batch service*: all particles present at a given service station are served collectively (and thus move together to the next service station, or out of the system). An occupancy-service classification of one-dimensional lattice-gas models is presented in Table I. We note that intermediate service capacities (larger than one but finite) are also possible [7,31].

The tandem Jackson network belongs to the wide family of Jackson networks, a fundamental class of queueing systems [28–30]. Devised in the early 1960s, Jackson networks were applied to model the then emerging packet-switched networks (e.g., the ARPANET), which turned out to be the precursors of the nowadays World Wide Web. Jackson networks (and related queueing networks) have also been studied in various physical contexts [1–4] and are closely related to the zero-range process [32–34].

In the zero-range process, particles hop from site to site on a network [usually a one-dimensional (1D) lattice] with a hop rate that depends, most generally, on the number of particles at the departure site and on the departure site itself. In the zero-range process, each site may be occupied by an arbitrary number of particles, and interactions are mediated via the dependence of the hopping rate on the number of particles that occupy a given site. The tandem Jackson network can be viewed as a zero-range process in which hopping rates depend only on the departure site (and are independent of the number of particles occupying this site) and general Jackson networks can be mapped onto the zero-range process as well. Interestingly, both Jackson networks and the zero-range process are characterized by steady-state distributions that are of a product form. This fact implies that the number of particles in a given site is statistically independent from the number of particles in any other site throughout the system. In light of this fact, the ASIP is very much distinct. While sharing resemblance with the tandem Jackson network and with the zero-range process, the steady-state distribution of the ASIP is not characterized by a simple product form. This fact is a mere manifestation of the fact that, in the ASIP, the occupancy of a given site depends on the occupancies of all the sites that precede it. Note, however, that the ASIP model is exactly solvable.

D. Batch service and growth-collapse processes

The notion of batch service is strongly related to growth-collapse processes, which play an important role in both queueing theory and statistical physics. Consider a single service station with batch service. Jobs arrive to the station randomly in time, causing the queue to grow steadily; when service is rendered, all jobs are served simultaneously, causing the queue to collapse to zero. Stochastic growth-collapse temporal patterns thus emerge from the application of batch-service policies [7,8,10,35,36]. In statistical physics, growth-collapse temporal patterns emerge in a host of complex systems, examples including sand-pile models and systems in self-organized criticality [37], stick-slip models of interfacial friction [38], Burridge-Knopof–type models of earthquakes and continental drift [39], stochastic Ornstein-Uhlenbeck capacitors [40], and geometric Langevin equations [41]. The ASIP model is, in effect, a tandem array of growth-collapse processes.

E. General outline of this paper

In this paper, we comprehensively explore the ASIP model. Our focus is set on the analysis of the stochastic dynamics and the stationary statistics of the ASIP's occupancy vector, the n-dimensional vector counting the number of particles present in each lattice site (at any given time). We derive evolution equations and steady-state equations for the mean and for the probability generating function (PGF) of the ASIP's occupancy vector. Explicit steady-state solutions are obtained for the mean. Explicit steady-state solutions are also obtained for the PGF of small ASIP systems ($n = 1,2,3$), and a computational scheme for solving the steady-state PGF equations for ASIP systems of arbitrary size is presented. We show that the steady-state PGF solutions explode in complexity as the lattice size increases, thus rendering the ASIP's occupancy vector analytically intractable for large n.

The ASIP's load is the total number of particles present in the lattice. In comparison to the ASIP's occupancy vector, analytical tractability of the ASIP's load is much more simple. Indeed, we obtain closed-form results for the mean, variance, and PGF of the ASIP's load in steady state. Interestingly, the load's PGF admits a product form representation, which, in turn, implies a surprising stochastic decomposition structure. Moreover, with the explicit steady-state load results at hand, we further study load optimization in steady state, seeking system parameters, namely, the rates of the underlying exponential clocks that optimize the ASIP in various aspects. Our analysis concludes that optimality is attained by *homogeneous* ASIP systems in which the underlying exponential clocks all have the same rate.

The remainder of the paper is organized as follows. In Sec. II, we define the ASIP, derive its Markovian law of motion,

and present a Monte Carlo algorithm for the simulation of its stochastic evolution. The mean analysis and the PGF analysis of the ASIP are carried out, respectively, in Secs. III and IV. The solution of the ASIP's PGF in steady state is discussed in Sec. V. The ASIP's load and load optimization are analyzed, respectively, in Secs. VI and VII.

II. ASIP MODEL

The asymmetric inclusion process is described as follows. Consider a system composed of a sequence of n gates labeled $k = 1, \ldots, n$. Each gate is preceded by a waiting zone, and the one preceding gate k is henceforth referred to as the kth waiting zone ($k = 1, \ldots, n$). Particles arrive at the first waiting zone following a Poisson process Π_0 with rate λ, the openings of gate k follow a Poisson process Π_k with rate μ_k, and the Poisson processes $\{\Pi_0, \Pi_1, \ldots, \Pi_n\}$ are independent. At an opening of gate k (gate-opening instant), all particles that are present at the kth waiting zone move on to the $(k+1)$th waiting zone, thus joining the particles already present in the latter waiting zone ($k = 1, \ldots, n-1$). At an opening of gate n, all particles that are present at the nth waiting zone leave the system. We emphasize that the very definition of the Poisson process precludes the possibility of two or more gates opening simultaneously (indeed such events have zero probability). Time durations between consecutive gate-opening events are exponentially distributed and, thus, series of hops can occur within an arbitrarily small time interval, but *can not* occur simultaneously. Particles' motion is hence restricted to jumps between consecutive sites. The ASIP model is illustrated in Fig. 1.

In the ASIP model, the capacity of the waiting zones is unbounded. Namely, the number of particles allowed to accumulate in each waiting zone is unlimited. In general, one can consider waiting zones with finite capacity, where l_k is the maximum capacity of the kth waiting zone ($k = 1, \ldots, n$; $l_k = 1, 2, \ldots$). In this case, at an opening of gate k, not all particles that are present at the kth waiting zone shall necessarily be allowed to move on to the $(k+1)$th waiting zone ($k = 1, \ldots, n-1$). Specifically, if at the opening instant of gate k there are x_k particles in the kth waiting zone, and x_{k+1} particles in the $(k+1)$th waiting zone, then exactly $\min(x_k, l_{k+1} - x_{k+1})$ particles will move from the kth waiting zone to the $(k+1)$th waiting zone. Note that the ASEP model is attained by the capacity limits $l_k = 1$, and that the ASIP model is attained by the capacity limits $l_k = \infty$ ($k = 1, \ldots, n$).

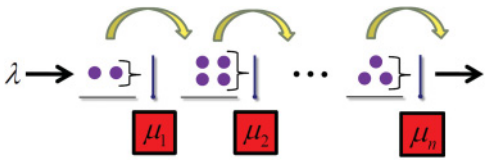

FIG. 1. (Color online) An illustration of the ASIP model. Particles arrive at the first waiting zone following a Poisson process with rate λ. The times between consecutive openings of gate k are exponentially distributed with mean $1/\mu_k$. The inflow process and the gate openings are assumed independent. At an opening of gate k, all particles present at the kth waiting zone move on to the $(k+1)$th waiting zone, thus joining the particles already present in the latter waiting zone ($k = 1, \ldots, n-1$). At an opening of gate n, all particles that are present at the nth waiting zone leave the system.

We now turn to describe the traversal time, the Markovian dynamics, and the Monte Carlo simulation of the ASIP model. Henceforth, we shall use the shorthand notation $\mu = \mu_1 + \cdots + \mu_n$ for the system's cumulative service rate.

A. Traversal time

Consider the random time T it takes a particle to traverse the system, henceforth termed the ASIP's *traversal time*. That is, T is the time elapsing from the instant a particle arrives at the first waiting zone until the instant it leaves the system. Due to the memoryless property of the exponential distribution, the time elapsing from the arrival of a particle to waiting zone k (at an arbitrary time epoch) until the first opening of gate k thereafter, is exponentially distributed with mean $1/\mu_k$. A particle arriving to the system would thus wait an exponentially distributed random time (with mean $1/\mu_1$) until moving from the first waiting zone to the second waiting zone, then wait an exponentially distributed random time (with mean $1/\mu_2$) until moving from the second waiting zone to the third waiting zone, and so forth. Since the gate openings are governed by independent Poisson processes, we conclude that the traversal time T admits the following stochastic representation:

$$T = \mathcal{E}_1 + \cdots + \mathcal{E}_n, \tag{1}$$

where $\{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$ is a sequence of independent and exponentially distributed random times with corresponding means $\{1/\mu_1, \ldots, 1/\mu_n\}$. Consequently, the mean and the variance of the traversal time T are given, respectively, by

$$\mathrm{E}[T] = \frac{1}{\mu_1} + \cdots + \frac{1}{\mu_n} \tag{2}$$

and

$$\mathrm{Var}[T] = \frac{1}{\mu_1^2} + \cdots + \frac{1}{\mu_n^2}. \tag{3}$$

B. Markovian dynamics

Let $X_k(t)$ denote the number of particles present in the kth waiting zone ($k = 1, \ldots, n$) at time t ($t \geqslant 0$), and set $\mathbf{X}(t) = [X_1(t), \ldots, X_n(t)]$. The vector $\mathbf{X}(t)$ represents the system's occupancy at time t. Observe the system at times t and $t' = t + \Delta$ (for small Δ) and use the shorthand notation $\mathbf{X} = \mathbf{X}(t)$ and $\mathbf{X}' = \mathbf{X}(t')$. The stochastic connection between the random vectors $\mathbf{X}$ and $\mathbf{X}'$, characterizing the Markovian law of motion

of the stochastic process $[\mathbf{X}(t)]_{t\geqslant 0}$, is given by

$$(X_1',\ldots,X_n') = \begin{cases} (X_1,X_2,X_3,\ldots,X_{n-1},X_n) & \text{w.p. } 1-(\lambda+\mu)\Delta+o(\Delta), \\ (X_1+1,X_2,X_3,\ldots,X_{n-1},X_n) & \text{w.p. } \lambda\Delta+o(\Delta), \\ (0,X_1+X_2,X_3,\ldots,X_{n-1},X_n) & \text{w.p. } \mu_1\Delta+o(\Delta), \\ (X_1,0,X_2+X_3,\ldots,X_{n-1},X_n) & \text{w.p. } \mu_2\Delta+o(\Delta), \\ \vdots & \vdots \\ (X_1,X_2,X_3,\ldots,0,X_{n-1}+X_n) & \text{w.p. } \mu_{n-1}\Delta+o(\Delta), \\ (X_1,X_2,X_3,\ldots,X_{n-1},0) & \text{w.p. } \mu_n\Delta+o(\Delta), \end{cases} \tag{4}$$

where w.p. stand for "with probability".

Equation (4) follows from considering the totality of events that may take place within the time interval $(t,t']$. There are $n+1$ such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first waiting zone, occurring with probability $\lambda\Delta+o(\Delta)$, in which case $X_1 \mapsto X_1' = X_1+1$; (1) opening of the first gate, occurring with probability $\mu_1\Delta+o(\Delta)$, in which case $X_1 \mapsto X_1' = 0$ and $X_2 \mapsto X_2' = X_1+X_2$; (2) opening of the second gate, occurring with probability $\mu_2\Delta+o(\Delta)$, in which case $X_2 \mapsto X_2' = 0$ and $X_3 \mapsto X_3' = X_2+X_3$; …; $(n-1)$ opening of the gate before last, occurring with probability $\mu_{n-1}\Delta+o(\Delta)$, in which case $X_{n-1} \mapsto X_{n-1}' = 0$ and $X_n \mapsto X_n' = X_{n-1}+X_n$; (n) opening of the last gate, occurring with probability $\mu_n\Delta+o(\Delta)$, in which case $X_n \mapsto X_n' = 0$. The first line on the right-hand side of Eq. (4) represents the scenario in which no event takes place, which occurs with the complementary probability $1-(\lambda+\mu)\Delta+o(\Delta)$.

C. Monte Carlo simulation

The ASIP's random trajectory $[\mathbf{X}(t)]_{t\geqslant 0}$ changes discretely rather than continuously. Indeed, between the underlying Poissonian events, arrival of a particle to the system, or an opening of one of the n gates, the ASIP's trajectory does not change. Consider now the ASIP's trajectory at the instant it changes (i.e., arrival of a particle or an opening of a gate). Let $Y_k(s)$ denote the number of particles present in the kth waiting zone ($k=1,\ldots,n$) immediately after the sth Poissonian event took place ($s=1,2,\ldots$), and set $\mathbf{Y}(s)=[Y_1(s),\ldots,Y_n(s)]$. Observe the system at two consecutive Poissonian events, s and $s'=s+1$, and use the shorthand notation $\mathbf{Y}=\mathbf{Y}(s)$ and $\mathbf{Y}'=\mathbf{Y}(s')$. The properties of the exponential distribution imply the following [42]:

(i) The time elapsing between two consecutive Poissonian events s and $s'=s+1$ is exponentially distributed with mean $1/(\lambda+\mu)$.

(ii) The stochastic connection between the random vectors $\mathbf{Y}$ and $\mathbf{Y}'$, characterizing the Markovian law of motion of the stochastic process $[(\mathbf{Y}(s)]_{s=1}^{\infty}$, is given by

$$(Y_1',\ldots,Y_n') = \begin{cases} (Y_1+1,Y_2,Y_3,\ldots,Y_{n-1},Y_n) & \text{w.p. } \lambda/(\lambda+\mu), \\ (0,Y_1+Y_2,Y_3,\ldots,Y_{n-1},Y_n) & \text{w.p. } \mu_1/(\lambda+\mu), \\ (Y_1,0,Y_2+Y_3,\ldots,Y_{n-1},Y_n) & \text{w.p. } \mu_2/(\lambda+\mu), \\ \vdots & \vdots \\ (Y_1,Y_2,Y_3,\ldots,0,Y_{n-1}+Y_n) & \text{w.p. } \mu_{n-1}/(\lambda+\mu), \\ (Y_1,Y_2,Y_3,\ldots,Y_{n-1},0) & \text{w.p. } \mu_n/(\lambda+\mu). \end{cases} \tag{5}$$

(iii) The time elapsing between the two consecutive Poissonian events s and $s'=s+1$, and the change $\mathbf{Y}\mapsto\mathbf{Y}'$, are mutually independent.

Equation (5) follows from considering the totality of events that lead to a change $\mathbf{Y}\mapsto\mathbf{Y}'$. There are $n+1$ such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first waiting zone,

occurring with probability $\lambda/(\lambda+\mu)$, in which case $Y_1 \mapsto Y_1' = Y_1 + 1$; (1) opening of the first gate, occurring with probability $\mu_1/(\lambda+\mu)$, in which case $Y_1 \mapsto Y_1' = 0$ and $Y_2 \mapsto Y_2' = Y_1 + Y_2$; (2) opening of the second gate, occurring with probability $\mu_2/(\lambda+\mu)$, in which case $Y_2 \mapsto Y_2' = 0$ and $Y_3 \mapsto Y_3' = Y_2 + Y_3$; ...; $(n-1)$ opening of the gate before last, occurring with probability $\mu_{n-1}/(\lambda+\mu)$, in which case $Y_{n-1} \mapsto Y_{n-1}' = 0$ and $Y_n \mapsto Y_n' = Y_{n-1} + Y_n$; (n) opening of the last gate, occurring with probability $\mu_n/(\lambda+\mu)$, in which case $Y_n \mapsto Y_n' = 0$. Properties (i), (ii), and (iii) establish, in effect, a simple and straightforward Monte Carlo algorithm for the simulation of the ASIP's random trajectory.

III. MEAN DYNAMICS AND MEAN FIELD ANALYSIS

In this section, we study the dynamics of the *means* of the random vectors $\mathbf{X}(t)$ $(t \geqslant 0)$ and $\mathbf{Y}(s)$ $(s = 1,2,\ldots)$. Throughout the section, we use the shorthand vector notation $\boldsymbol{\lambda} = (\lambda,0,\ldots,0)^T$, and the shorthand matrix notation

$$\mathbf{M} = \begin{pmatrix} -\mu_1 & & & \cdots & & \\ \mu_1 & -\mu_2 & & \cdots & & \\ & \mu_2 & -\mu_3 & & & \\ & & \ddots & \ddots & & \\ & & & \mu_{n-2} & -\mu_{n-1} & \\ & & & & \mu_{n-1} & -\mu_n \end{pmatrix}. \tag{6}$$

(Note: In the above matrix, all blank spaces represent zero entries.)

A. Mean dynamics of $\mathbf{X}(t)$

We denote the mean of the random vector $\mathbf{X}(t)$ by

$$\mathbf{e}_{\mathbf{X}}(t) = (\mathrm{E}[X_1(t)],\ldots,\mathrm{E}[X_n(t)])^{\top}. \tag{7}$$

Observe the system at times t and $t' = t + \Delta$. Conditioning on $\mathbf{X}(t)$ and utilizing the Markovian dynamics of Eq. (4) yields

$$\mathrm{E}[\mathbf{X}(t')] = \mathrm{E}[\mathrm{E}[\mathbf{X}(t')|\,\mathbf{X}(t)]] = \begin{cases} (1-(\lambda+\mu)\Delta)\mathrm{E}[\mathbf{X}(t)] + \lambda\Delta\mathrm{E}[\mathbf{X}(t)+(1,0,\ldots,0)^{\top}] \\ + \\ (\mu_1\Delta)\mathrm{E}[\mathbf{X}(t)+[-X_1(t),X_1(t),0,\ldots,0]^{\top}] \\ + \\ (\mu_2\Delta)\mathrm{E}[\mathbf{X}(t)+[0,-X_2(t),X_2(t),\ldots,0]^{\top}] \\ +\cdots+ \\ (\mu_{n-1}\Delta)\mathrm{E}[\mathbf{X}(t)+[0,\ldots,0,-X_{n-1}(t),X_{n-1}(t)]^{\top}] \\ + \\ (\mu_n\Delta)\mathrm{E}[\mathbf{X}(t)+[0,\ldots,0,-X_n(t)]^{\top}] \\ + \\ o(\Delta). \end{cases} \tag{8}$$

By rearranging the terms of Eq. (8), dividing by Δ, and taking $\Delta \to 0$, we conclude that

$$\frac{d\mathbf{e}_{\mathbf{X}}}{dt}(t) = \mathbf{M}\mathbf{e}_{\mathbf{X}} + \boldsymbol{\lambda}. \tag{9}$$

Equation (9) represents the mean dynamics of the random vector $\mathbf{X}(t)$. Namely, it transforms the Markovian dynamics of Eq. (4) to a differential equation that governs the temporal evolution of the mean vector $\mathbf{e}_{\mathbf{X}}(t)$ $(t \geqslant 0)$. The solution of Eq. (9) can be shown to be given by

$$\mathbf{e}_{\mathbf{X}}(t) = \mathbf{M}^{-1}[\exp(\mathbf{M}t) - \mathbf{I}]\boldsymbol{\lambda}. \tag{10}$$

B. Mean dynamics of $\mathbf{Y}(s)$

We denote the mean of the random vector $\mathbf{Y}(s)$ by

$$\mathbf{e}_{\mathbf{Y}}(s) = (\mathrm{E}[Y_1(s)],\ldots,\mathrm{E}[Y_n(s)])^{\top}. \tag{11}$$

Observe the system at two consecutive s and $s' = s+1$ Poissonian events. Conditioning on $\mathbf{Y}(s)$ and utilizing the law of motion presented in Eq. (5) yields

$$\begin{aligned} \mathrm{E}[\mathbf{Y}(s')] &= \mathrm{E}[\mathrm{E}[\mathbf{Y}(s')|\mathbf{Y}(s)]] \\ &= \begin{cases} \frac{\lambda}{\lambda+\mu}\mathrm{E}[\mathbf{Y}(s)+(1,0,\ldots,0)^{\top}] \\ + \\ \frac{\mu_1}{\lambda+\mu}\mathrm{E}[\mathbf{Y}(s)+[-Y_1(s),Y_1(s),0,\ldots,0]^{\top}] \\ + \\ \frac{\mu_2}{\lambda+\mu}\mathrm{E}[\mathbf{Y}(s)+[0,-Y_2(s),Y_2(s),\ldots,0]^{\top}] \\ +\cdots+ \\ \frac{\mu_{n-1}}{\lambda+\mu}\mathrm{E}[\mathbf{Y}(s)+[0,\ldots,0,-Y_{n-1}(s),Y_{n-1}(s)]^{\top}] \\ + \\ \frac{\mu_n}{\lambda+\mu}\mathrm{E}[\mathbf{Y}(s)+[0,\ldots,0,-Y_n(s)]^{\top}]. \end{cases} \end{aligned} \tag{12}$$

Rearranging the terms of Eq. (12), we conclude that

$$(\lambda+\mu)[\mathbf{e}_{\mathbf{Y}}(s') - \mathbf{e}_{\mathbf{Y}}(s)] = \mathbf{M}\mathbf{e}_{\mathbf{Y}}(s) + \boldsymbol{\lambda}. \tag{13}$$

Equation (13) represents the mean dynamics of the random vector $\mathbf{Y}(s)$. That is, it transforms the law of motion of Eq. (5) to a difference equation that governs the temporal evolution of the mean vector $\mathbf{e_Y}(s)$ $(s = 1,2,\ldots)$. The solution of Eq. (13) can be shown to be given by

$$\mathbf{e_Y}(s) = \mathbf{M}^{-1}\left[\left(\mathbf{I} + \frac{1}{\lambda+\mu}\mathbf{M}\right)^{s} - \mathbf{I}\right]\boldsymbol{\lambda}. \tag{14}$$

C. Mean field analysis in steady state

Consider now the ASIP model in *steady state*. In steady state, the stochastic processes $[\mathbf{X}(t)]_{t\geqslant 0}$ and $[\mathbf{Y}(s)]_{s=1}^{\infty}$ are *stationary*, and hence their respective means are time homogeneous: $\mathbf{e_X}(t) \equiv \mathbf{e_X}$ $(t \geqslant 0)$ and $\mathbf{e_Y}(s) \equiv \mathbf{e_Y}$ $(s = 1,2,\ldots)$. Substituting the time-homogeneous vectors $\mathbf{e_X}(t) \equiv \mathbf{e_X}$ and $\mathbf{e_Y}(s) \equiv \mathbf{e_Y}$, respectively, into Eqs. (9) and (13), yields the common equation

$$0 = \mathbf{M}\mathbf{e} + \boldsymbol{\lambda} \tag{15}$$

[where $\mathbf{e} = (e_1,\ldots,e_n)$ is the *unknown* vector]. Namely, both the mean vectors $\mathbf{e_X}$ and $\mathbf{e_Y}$ are governed by Eq. (15).

A straightforward computation of Eq. (15) yields the steady-state solution

$$e_k = \mathrm{E}[X_k(t)] = \mathrm{E}[Y_k(s)] = \frac{\lambda}{\mu_k} \tag{16}$$

$(k = 1,\ldots,n)$. Combining Eqs. (2) and (16) together further yields the following steady-state formula:

$$\mathrm{E}\left[\sum_{k=1}^{n} X_k(t)\right] = \mathrm{E}\left[\sum_{k=1}^{n} Y_k(s)\right] = \sum_{k=1}^{n}\frac{\lambda}{\mu_k} = \lambda\mathrm{E}[T]. \tag{17}$$

Equation (17) asserts that, at steady state, the mean number of particles in the system is given by the product $\lambda\mathrm{E}[T]$: the flow rate λ into the system times the mean traversal time $\mathrm{E}[T]$, the mean sojourn time of an arbitrary particle in the system. Note that, although the random variables $\{X_1(t),\ldots,X_n(t)\}$ [and similarly $\{Y_1(s),\ldots,Y_n(s)\}$] are intricately dependent, these dependencies do not affect the mean behavior given by Eq. (17). Equation (17) is the ASIP version of the well known Little's law in queueing theory [43].

D. Beyond the mean field description

In the ASEP model, the mean approximation well describes the system: the ASEP statistical behavior can be represented by a mean field plus an additional small noise term. Moreover, the ASEP's mean field approximation improves as the system size grows larger. The statistical behavior of the ASIP model is dramatically different: fluctuations of the ASIP's occupancy vector grow as the system becomes larger. This phenomenon is demonstrated in Fig. 2, in which a homogeneous ASIP system is simulated: For each site k, we numerically calculate the steady-state mean and standard deviation of the number of particles X_k present in the kth waiting zone; the simulation vividly shows a *power law* growth of the standard deviation as a function of the system's size (number of gates). The dramatic difference between the statistical behaviors of the ASEP and ASIP models is due to the different "service policies" of these models, as expressed in Table I. Contrary to the ASEP model, the batch service of the ASIP model causes an "all or none" effect, which, in turn, leads to site occupancy fluctuations, the typical order of magnitude of which can be much larger than the mean occupancy itself.

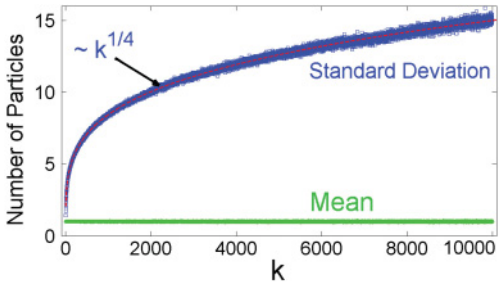

FIG. 2. (Color online) Large fluctuations and the emergence of scaling laws in the ASIP model. We have simulated a homogeneous ASIP system with 10 000 sites, $\lambda = 1$, and $\mu_k = 1$ for $k = 1,\ldots,10\,000$. The mean and standard deviation in the number of particles at site k are plotted as a function of the site index. As expected, we find that, regardless of the site index, on average, each site is occupied by a single particle ($e_k = \frac{\lambda}{\mu_k} = 1$). Conversely, the standard deviation in the number of particles exhibits a power law dependence on k and grows like $\sim k^{1/4}$ [dashed line is given by $N(k) = 1.5 \times k^{1/4}$]. As fluctuations around the mean are typically much larger than the mean itself, it is clear that a "mean field" description is unable to capture the physics of large ASIP systems.

The mean field approximation fails to capture the complexity and the interdependencies of the ASIP model. To fully capture the intricate statistics of the ASIP model, we now turn to analyze its multidimensional probability distributions via probability generating functions.

IV. PGF DYNAMICS

In this section, we study the dynamics of the *probability generating functions* (PGFs) of the random vectors $\mathbf{X}(t)$ $(t \geqslant 0)$ and $\mathbf{Y}(s)$ $(s = 1,2,\ldots)$.

A. PGF dynamics of X(t)

The PGF of the random vector $\mathbf{X}(t)$ is given by

$$G_{\mathbf{X}}(t,z_1,z_2,\ldots,z_n) = \mathrm{E}\left[z_1^{X_1(t)}z_2^{X_2(t)},\ldots,z_n^{X_n(t)}\right] \tag{18}$$

$(|z_k| \leqslant 1,\ k = 1,\ldots,n)$. Observe the system at times t and $t' = t + \Delta$ and use again the shorthand notation $\mathbf{X} = \mathbf{X}(t)$ and $\mathbf{X}' = \mathbf{X}(t')$. By conditioning on $\mathbf{X}$ and utilizing the Markovian

dynamics of Eq. (4), we have

$$\mathrm{E}\left[\prod_{k=1}^{n} z_k^{X_k'}\right] = \mathrm{E}\left[\mathrm{E}\left[\prod_{k=1}^{n} z_k^{X_k'} | \mathbf{X}\right]\right]$$
$$= \begin{cases} [1-(\lambda+\mu)\Delta]\mathrm{E}\left[\prod_{k=1}^{n} z_k^{X_k}\right] + (\lambda\Delta)\mathrm{E}\left[z_1 \prod_{k=1}^{n} z_k^{X_k}\right] \\ + \\ (\mu_1\Delta)\mathrm{E}\left[z_2^{X_1} \prod_{k\neq 1}^{n} z_k^{X_k}\right] + (\mu_2\Delta)\mathrm{E}\left[z_3^{X_2} \prod_{k\neq 2}^{n} z_k^{X_k}\right] \\ + \cdots + \\ (\mu_{n-1}\Delta)\mathrm{E}\left[z_n^{X_{n-1}} \prod_{k\neq (n-1)}^{n} z_k^{X_k}\right] + (\mu_n\Delta)\mathrm{E}\left[\prod_{k\neq n}^{n} z_k^{X_k}\right] \\ + \\ o(\Delta). \end{cases} \tag{19}$$

Using the PGF notation of Eq. (18), Eq. (19) reads as

$$G_{\mathbf{X}}(t', z_1, z_2, \ldots, z_n)$$
$$= \begin{cases} [1-(\lambda+\mu)\Delta]G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_{n-1}, z_n) \\ + \\ (\lambda\Delta) z_1 G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_{n-1}, z_n) \\ + \\ (\mu_1\Delta) G_{\mathbf{X}}(t, z_2, z_2, z_3, \ldots, z_{n-1}, z_n) \\ + \\ (\mu_2\Delta) G_{\mathbf{X}}(t, z_1, z_3, z_3, \ldots, z_{n-1}, z_n) \\ + \cdots + \\ (\mu_{n-1}\Delta) G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_n, z_n) \\ + \\ (\mu_n\Delta) G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_{n-1}, 1) \\ + \\ o(\Delta). \end{cases} \tag{20}$$

By rearranging the terms of Eq. (20), dividing by Δ, and taking $\Delta \to 0$, we conclude that

$$\frac{\partial G_{\mathbf{X}}}{\partial t}(t, z_1, \ldots, z_n)$$
$$= \begin{cases} [\lambda(z_1 - 1) - \mu]G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_{n-1}, z_n) \\ + \\ \mu_1 G_{\mathbf{X}}(t, z_2, z_2, z_3, \ldots, z_{n-1}, z_n) \\ + \\ \mu_2 G_{\mathbf{X}}(t, z_1, z_3, z_3, \ldots, z_{n-1}, z_n) \\ + \cdots + \\ \mu_{n-1} G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_n, z_n) \\ + \\ \mu_n G_{\mathbf{X}}(t, z_1, z_2, z_3, \ldots, z_{n-1}, 1). \end{cases} \tag{21}$$

Equation (21) represents the PGF dynamics of the random vector $\mathbf{X}(t)$. Namely, it transforms the Markovian dynamics of (4) to a differential equation of the form

$$\frac{\partial G_{\mathbf{X}}}{\partial t}(t, \mathbf{z}) = [\mathcal{A} G_{\mathbf{X}}](t, \mathbf{z}), \tag{22}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, and where $\mathcal{A}$ is an operator that acts only on the $\mathbf{z}$ part of the PGF $G_{\mathbf{X}}(t, \mathbf{z})$.

B. PGF dynamics of $\mathbf{Y}(s)$

The PGF of the random vector $\mathbf{Y}(s)$ is given by

$$G_{\mathbf{Y}}(s, z_1, z_2, \ldots, z_n) = \mathbf{E}\left[z_1^{Y_1(s)} z_2^{Y_2(s)}, \ldots, z_n^{Y_n(s)}\right] \tag{23}$$

($|z_k| \leqslant 1$, $k = 1, \ldots, n$). Observe the system at two consecutive s and $s' = s + 1$ Poissonian events, and use again the shorthand notation $\mathbf{Y} = \mathbf{Y}(s)$ and $\mathbf{Y}' = \mathbf{Y}(s')$. By conditioning on $\mathbf{Y}$ and utilizing the law of motion of Eq. (5), we have

$$\mathrm{E}\left[\prod_{k=1}^{n} z_k^{Y_k'}\right] = \mathrm{E}\left[\mathrm{E}\left[\prod_{k=1}^{n} z_k^{Y_k'} | \mathbf{Y}\right]\right]$$
$$= \begin{cases} \frac{\lambda}{\lambda+\mu}\mathrm{E}\left[z_1 \prod_{k=1}^{n} z_k^{Y_k}\right] + \frac{\mu_1}{\lambda+\mu}\mathrm{E}\left[z_2^{Y_1} \prod_{k\neq 1}^{n} z_k^{Y_k}\right] \\ + \\ \frac{\mu_2}{\lambda+\mu}\mathrm{E}\left[z_3^{Y_2} \prod_{k\neq 2}^{n} z_k^{Y_k}\right] + \frac{\mu_3}{\lambda+\mu}\mathrm{E}\left[z_4^{Y_3} \prod_{k\neq 3}^{n} z_k^{Y_k}\right] \\ + \cdots + \\ \frac{\mu_{n-1}}{\lambda+\mu}\mathrm{E}\left[z_n^{Y_{n-1}} \prod_{k\neq (n-1)}^{n} z_k^{Y_k}\right] + \frac{\mu_n}{\lambda+\mu}\mathrm{E}\left[\prod_{k\neq n}^{n} z_k^{Y_k}\right]. \end{cases} \tag{24}$$

By using the PGF notation Eq. (23) and rearranging terms, Eq. (24) reads as

$$G_{\mathbf{Y}}(s', z_1, \ldots, z_n) - G_{\mathbf{Y}}(s, z_1, \ldots, z_n)$$
$$= \begin{cases} \frac{\lambda(z_1-1)-\mu}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \ldots, z_n) \\ + \\ \frac{\mu_1}{\lambda+\mu} G_{\mathbf{Y}}(s, z_2, z_2, \ldots, z_n) \\ + \\ \frac{\mu_2}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, z_3, z_3, \ldots, z_n) \\ + \cdots + \\ \frac{\mu_{n-1}}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \ldots, z_n, z_n) \\ + \\ \frac{\mu_n}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \ldots, z_{n-1}, 1). \end{cases} \tag{25}$$

Equation (25) represents the PGF dynamics of the random vector $\mathbf{Y}(s)$. Namely, it transforms the law of motion of Eq. (25) to a difference equation of the form

$$G_{\mathbf{Y}}(s', \mathbf{z}) - G_{\mathbf{Y}}(s, \mathbf{z}) = [\mathcal{B} G_{\mathbf{Y}}](s, \mathbf{z}), \tag{26}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, and where $\mathcal{B}$ is an operator that acts only the $\mathbf{z}$ part of the PGF $G_{\mathbf{Y}}(s, \mathbf{z})$.

C. Steady state

Consider now the ASIP model in *steady state*. The stochastic processes $[\mathbf{X}(t)]_{t\geqslant 0}$ and $[\mathbf{Y}(s)]_{s=1}^{\infty}$ are *stationary* and, hence, their respective PGFs are time homogeneous: $G_{\mathbf{X}}(t, \mathbf{z}) \equiv G_{\mathbf{X}}(\mathbf{z})$ ($t \geqslant 0$) and $G_{\mathbf{Y}}(s, \mathbf{z}) \equiv G_{\mathbf{Y}}(\mathbf{z})$ ($s = 1, 2, \ldots$). Substituting the time-homogeneous PGFs $G_{\mathbf{X}}(t, \mathbf{z}) \equiv G_{\mathbf{X}}(\mathbf{z})$ and $G_{\mathbf{Y}}(s, \mathbf{z}) \equiv G_{\mathbf{Y}}(\mathbf{z})$, respectively, into Eqs. (21) and (25) yields

the common equation

$$[\lambda(1-z_1)+\mu]G(z_1,z_2,z_3,\ldots,z_{n-1},z_n) = \begin{cases} \mu_1 G(z_2,z_2,z_3,\ldots,z_{n-1},z_n) \\ + \\ \mu_2 G(z_1,z_3,z_3,\ldots,z_{n-1},z_n) \\ +\cdots+ \\ \mu_{n-1} G(z_1,z_2,z_3,\ldots,z_n,z_n) \\ + \\ \mu_n G(z_1,z_2,z_3,\ldots,z_{n-1},1) \end{cases} \tag{27}$$

[where $G(\mathbf{z})$ is the unknown function]. Namely, both the PGFs $G_{\mathbf{X}}(\mathbf{z})$ and $G_{\mathbf{Y}}(\mathbf{z})$ are governed by Eq. (27).

Assuming that Eq. (27) admits a unique solution (we shall address the issue of uniqueness in Sec. V), we obtain the following: *In steady state, the distribution of the vector* $\mathbf{X}(t)$ *coincides with the distribution of the vector* $\mathbf{Y}(s)$. Namely, in steady state, the ASIP displays the same statistics at arbitrary time and Poissonian events time epochs. In the nomenclature of queueing theory, such a phenomenon is termed PASTA (Poisson arrivals see time average [43]).

The PASTA phenomenon is a central concept in queueing theory, which implies that arriving customers find, on average, the same workload in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state S is exactly the same as the fraction of time the system is in state S. While well known results in queueing theory assert that the PASTA phenomenon holds for classes of systems with Poissonian arrivals (also known as $M/\cdot/\cdot$ queueing systems), this phenomenon does *not* hold for general systems. Indeed, even very simple queueing systems may fail to satisfy the PASTA phenomenon.

As an example, consider the $D/D/1$ queueing system. In this system, customers arrive to a service station with a single server in which they are processed according to their order of arrival. The customers' interarrival times and service times are deterministic. Let d_{arr} and d_{ser} denote, respectively, the deterministic interarrival and service times. Exactly every d_{arr} time units a new customer arrives at the service station, this customer must be served for exactly d_{ser} time units before leaving the system. Clearly, the queue will explode if $d_{\text{ser}} > d_{\text{arr}}$, will be perfectly balanced if $d_{\text{ser}} = d_{\text{arr}}$, and will be stationary if $d_{\text{ser}} \leqslant d_{\text{arr}}$. If $d_{\text{ser}} < d_{\text{arr}}$, then the queue cycles will coincide with the customers' arrival epochs, the server will be busy for d_{ser} time units after arrival, and will be idle in the remaining d_{arr}-d_{ser} time units. Clearly, arriving customers always observe an empty system (upon arrival). Hence, the fraction of customers finding the system nonempty is zero. On the other hand, the fraction of time the system is nonempty is $d_{\text{ser}}/d_{\text{arr}}$. The $DD1$ queueing model vividly exemplifies how the PASTA phenomenon can be violated even in very simple systems. On the other hand, the PASTA phenomenon can hold in complex processes such as the running maxima of nonlinear shot noise [44]. The fact that the PASTA phenomenon holds for all ASIP systems is far from being trivial.

We now turn to describe the embedding phenomenon, another useful property of the ASIP. Consider two ASIP models: model (A) with m gates and parameters $\{\lambda,\mu_1,\ldots,\mu_m\}$, and model (B) with n gates and parameters $\{\lambda,\mu_1,\ldots,\mu_n\}$, where $m < n$. Equation (27) implies the following embedding phenomenon: *The steady-state distribution of model (A) coincides with the steady-state distribution of the first* m *coordinates of model (B)*. The derivation of the embedding phenomenon follows from substituting $z_{m+1} = \cdots = z_n = 1$ in Eq. (27). The intuitive understanding of the embedding phenomenon follows from the fact that, in an ASIP model with n gates, the operation of the first m gates ($k = 1,\ldots,m$) is indifferent to whatever happens in the following gates ($k = m+1,\ldots,n$). In other words, an observation of the first m gates in an ASIP model with n gates is indistinguishable from an observation of an ASIP model with m gates (and the same parameters).

V. STEADY-STATE ANALYSIS

In this section, we explore Eq. (27) governing the steady-state PGF of the ASIP model.

A. Explicit solution: $n = 1$

Consider the ASIP model with a single gate ($n = 1$). In this case, Eq. (27) reduces to

$$[\lambda(1-z_1)+\mu_1]G(z_1) = \mu_1 G(1). \tag{28}$$

Noting that $G(1) = 1$, and setting $p_1 = \mu_1/(\mu_1+\lambda)$, Eq. (28) implies that

$$G(z_1) = \frac{\mu_1}{\lambda(1-z_1)+\mu_1} = \frac{p_1}{1-(1-p_1)z_1}. \tag{29}$$

The PGF of Eq. (29) characterizes the *geometric law* on the non-negative integers. Indeed, expanding both sides of Eq. (29) to power series (in the variable z_1) yields the probability distribution

$$\Pr(X_1 = j) = \Pr(Y_1 = j) = (1-p_1)^j p_1, \tag{30}$$

where $j = 0,1,2,\ldots$.

The probabilistic explanation of Eq. (30) is as follows: When $n = 1$, we can think about a competition between two Poissonian processes: gate openings and particle arrivals. The Poissonian nature of these processes implies that the probability that the first Poissonian event is an arrival of a particle is $1-p_1$. Similarly, the probability that the first Poissonian event is a gate opening is p_1. The memoryless property of the exponential distribution implies that, in order for exactly k particles to leave the system at a gate-opening moment, exactly k consecutive arrivals must be followed by a single gate opening. Hence, the random variable Y_1 is geometrically distributed (on the non-negative integers) with parameter p_1. As a result of the PASTA phenomenon described in the preceding section, the distribution of the system vector at steady state is equal in law to the distribution of the system vector immediately after Poissonian events, implying that X_1 coincides, in law, with Y_1.

B. Explicit solution: $n = 2$

In this section, we present, via the special case of $n = 2$, an iterative scheme for the solution of Eq. (27). In the basic step of the scheme, one uses Eq. (27) in order to obtain expressions for each of the generating functions that appear on its right-hand side. By repeating the basic step, time and again, a branching tree structure of generating functions forms. In this tree, each "parent" generating function is expressed by a set of "daughter" generating functions. As we shall demonstrate, the daughter generating functions become somewhat simpler with every step of the scheme. Eventually, the daughter generating functions become trivial, forming the "leaves" of our branching tree. The scheme terminates once all daughter generating functions turn into leaves. The PGF is then obtained from transcending upward from the leaves of the tree to its root. At the root, an explicit, and by construction *unique*, expression for the PGF is attained.

Consider the ASIP model with two gates ($n = 2$). In this case, Eq. (27) reduces to

$$[\lambda(1-z_1)+\mu_1+\mu_2]G(z_1,z_2) = \begin{cases} \mu_1 G(z_2,z_2) \\ + \\ \mu_2 G(z_1,1). \end{cases} \tag{31}$$

Now, following the scheme's basic step, we iteratively apply Eq. (31) to the daughters $G(z_2,z_2)$ and $G(z_1,1)$.

For the daughter $G(z_2,z_2)$, the basic step yields

$$[\lambda(1-z_2)+\mu_1+\mu_2]G(z_2,z_2) = \begin{cases} \mu_1 G(z_2,z_2) \\ + \\ \mu_2 G(z_2,1) \end{cases} \tag{32}$$

from which we obtain that

$$G(z_2,z_2) = \frac{\mu_2}{\lambda(1-z_2)+\mu_2} G(z_2,1). \tag{33}$$

In turn, for the daughter $G(z_2,1)$, Eq. (31) yields

$$[\lambda(1-z_2)+\mu_1+\mu_2]G(z_2,1) = \begin{cases} \mu_1 G(1,1) \\ + \\ \mu_2 G(z_2,1) \end{cases} \tag{34}$$

from which we obtain that

$$G(z_2,1) = \frac{\mu_1}{\lambda(1-z_2)+\mu_1} G(1,1). \tag{35}$$

For the daughter $G(z_1,1)$, the iteration yields

$$[\lambda(1-z_1)+\mu_1+\mu_2]G(z_1,1) = \begin{cases} \mu_1 G(1,1) \\ + \\ \mu_2 G(z_1,1) \end{cases} \tag{36}$$

from which we obtain that

$$G(z_1,1) = \frac{\mu_1}{\lambda(1-z_1)+\mu_1} G(1,1). \tag{37}$$

The leaves of our tree are characterized by the PGF $G(1,1)$, which trivially equals unity. Hence, setting $G(1,1) = 1$ in the leaves Eqs. (35) and (37) yields the daughters $G(z_2,1)$ and $G(z_1,1)$. Substituting the daughter $G(z_2,1)$ into Eq. (33) yields the daughter

$$G(z_2,z_2) = \frac{\mu_2}{\lambda(1-z_2)+\mu_2} \frac{\mu_1}{\lambda(1-z_2)+\mu_1}. \tag{38}$$

Finally, substituting the daughters $G(z_2,z_2)$ and $G(z_1,1)$ into Eq. (31) yields the root

$$G(z_1,z_2) = \begin{cases} \frac{\mu_1^2\mu_2}{[\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \\ + \\ \frac{\mu_1\mu_2}{[\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]}. \end{cases} \tag{39}$$

Summarizing, we have found that, for $n = 2$, the scheme terminates after two iterations. The result is a treelike structure, the leaves of which are trivial constants all equal to unity. Knowing the constants that stand in the base of the tree, we are able to calculate the functions that occupy the second lowest level. The PGF $G(z_1,z_2)$ was computed by iterating this procedure, i.e., by using known functions at the current knowledge level of the tree in order to compute the functions at the next level. A tree sketch of the solution steps for the ASIP model with two gates ($n = 2$) is depicted in Fig. 3.

C. Explicit solution: $n = 3$

The iterative scheme described in the preceding section applies, in theory, to the ASIP model with an arbitrary number of gates. In practice, however, the solution's complexity increases rapidly with the number of gates n. Thus, effectively, for large n, the PGF of the ASIP model is not tractable. To illustrate just how dramatically the solution complexity increases, consider the ASIP model with three gates ($n = 3$). In the Appendix, we show that the expression for $G(z_1,z_2,z_3)$

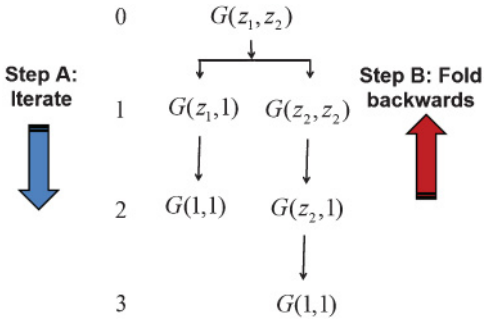

FIG. 3. (Color online) Schematic illustration of the iterative solution of Eq. (27) for $n = 2$. Step A: Eq. (27) is iterated repeatedly, in a branching-tree structure, until reaching the leaves $G(1,1) = 1$. Step B: The tree is folded back, yielding the value of the root $G(z_1,z_2)$ [Eq. (39)].

is given by

$$G(z_1,z_2,z_3)=\begin{cases}\dfrac{\mu_1^3\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_3)+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1]}\\ +\\ \dfrac{\mu_1^3\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1][\lambda(1-z_2)+\mu_1+\mu_2]}\\ +\\ \dfrac{\mu_1^2\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_1][\lambda(1-z_2)+\mu_1+\mu_2]}\\ +\\ \dfrac{\mu_1^2\mu_2\mu_3}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1]}\\ +\\ \dfrac{\mu_1^2\mu_2^2\mu_3}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_3)+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1]}\\ +\\ \dfrac{\mu_1^2\mu_2^2\mu_3/[\lambda(1-z_1)+\mu_1+\mu_3]}{[\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]}\\ +\\ \dfrac{\mu_1\mu_2^2\mu_3}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]}\\ +\\ \dfrac{\mu_1^2\mu_2\mu_3}{[\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]}\\ +\\ \dfrac{\mu_1\mu_2\mu_3}{[\lambda(1-z_1)+\mu][\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]}.\end{cases}\tag{40}$$

Equation (40) well exemplifies the intrinsic complexity of the ASIP model. A tree sketch of the solution steps for the ASIP model with three gates ($n=3$) is depicted in Fig. 4.

We note that at first glance it might seem possible to derive the steady-state *marginal* distributions of the random variables $\{X_1(t),\ldots,X_n(t)\}$ iteratively, namely, to establish a recursion equation relating the PGF of $X_k(t)$ to the PGF of $X_{k-1}(t)$, and then solve it. However, the random variables $\{X_1(t),\ldots,X_k(t)\}$ are correlated, and when trying to establish the aforementioned recursion equation for the PGF of $X_k(t)$, we end up with the multidimensional PGF of the entire vector $[X_1(t),\ldots,X_k(t)]$. This is yet another feature of the intractability of the ASIP model.

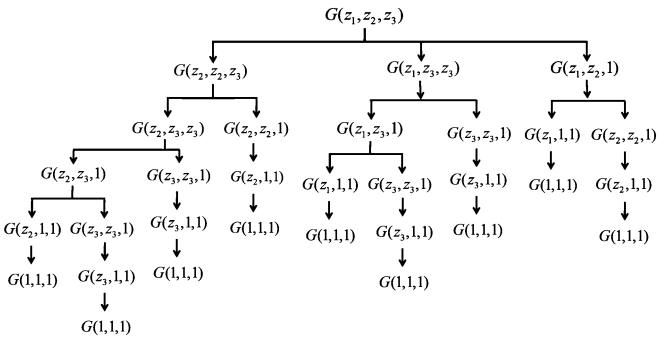

FIG. 4. Schematic illustration of the iterative solution of Eq. (27) for $n=3$. Step A: Eq. (27) is iterated repeatedly, in a branching-tree structure, until reaching the leaves $G(1,1,1)=1$. Step B: The tree is folded back, yielding the value of the root $G(z_1,z_2,z_3)$ [Eq. (40)].

VI. LOAD ANALYSIS

In Secs. IV and V, we studied the PGF dynamics and steady-state solution of the random vectors $\mathbf{X}(t)$ and $\mathbf{Y}(s)$, and demonstrated intractability for large n. In this section, we study the *load* of the ASIP model: the total number of particles in the system, given by the sum of coordinates of the random vectors $\mathbf{X}(t)$ and $\mathbf{Y}(s)$. Surprisingly so, the load turns out to be analytically tractable.

A. Load analysis of $\mathbf{X}(t)$

Let $X_{(k)}(t)$ denote the total number of particles present, at an arbitrary time t ($t\geqslant 0$), in the first k waiting zones. The random variable $X_{(k)}(t)$ is the sum of the first k coordinates of the random vector $\mathbf{X}(t)$, i.e., $X_{(k)}(t)=X_1(t)+\cdots+X_k(t)$ ($k=1,\ldots,n$).

Observe the system at times t and $t'=t+\Delta$, and use the shorthand notation $X_{(k)}=X_{(k)}(t)$ and $X'_{(k)}=X_{(k)}(t')$ ($k=1,\ldots,n$). Equation (4) implies that the stochastic connection between the random sums $X_{(k)}$ and $X'_{(k)}$, characterizing the Markovian dynamics of the stochastic process $[X_{(k)}(t)]_{t\geqslant 0}$, is given by

$$X'_{(k)}=\begin{cases}X_{(k)} & \text{w.p. } 1-(\lambda+\mu_k)\Delta+o(\Delta),\\ X_{(k)}+1 & \text{w.p. } \lambda\Delta+o(\Delta),\\ X_{(k-1)} & \text{w.p. } \mu_k\Delta+o(\Delta).\end{cases}\tag{41}$$

Equation (41) follows from considering the events that may take place and result in a change $X_{(k)}\mapsto X'_{(k)}\neq X_{(k)}$ within the time interval $(t,t']$. There are exactly two such events, and

we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first waiting zone, occurring with probability $\lambda\Delta + o(\Delta)$, in which case $X_{(k)} \mapsto X'_{(k)} = X_{(k)} + 1$; ($k$) opening of the kth gate, occurring with probability $\mu_k\Delta + o(\Delta)$, in which case $X_{(k)} \mapsto X'_{(k)} = X_{(k-1)}$. The first line on the right-hand side of Eq. (41) represents the scenario in which no event takes place, which occurs with the complementary probability $1 - (\lambda + \mu_k)\Delta + o(\Delta)$.

Let

$$G_{X_{(k)}}(t,z) = \mathrm{E}[z^{X_{(k)}(t)}] \tag{42}$$

($|z| \leqslant 1$) denote the PGF of the random sum $X_{(k)}(t)$. Setting $z_1 = \cdots = z_k = z$ and $z_{k+1} = \cdots = z_n = 1$, and noting that, by definition, $G_{X_{(k)}}(t,z) = G_{\mathbf{X}}(t,z,\ldots,z,1,\ldots,1)$, Eq. (21) yields

$$\frac{\partial G_{X_{(k)}}}{\partial t}(t,z) = \begin{cases} [\lambda(z-1) - \mu]G_{X_{(k)}}(t,z) \\ + \\ \mu_1 G_{X_{(k)}}(t,z) \\ + \cdots + \\ \mu_{k-1} G_{X_{(k)}}(t,z) \\ + \\ \mu_k G_{X_{(k-1)}}(t,z) \\ + \\ \mu_{k+1} G_{X_{(k)}}(t,z) \\ + \cdots + \\ \mu_{n-1} G_{X_{(k)}}(t,z) \\ + \\ \mu_n G_{X_{(k)}}(t,z). \end{cases} \tag{43}$$

Equation (43), in turn, implies that the PGF dynamics of the random sum $X_{(k)}(t)$ is given by

$$\frac{\partial G_{X_{(k)}}}{\partial t}(t,z) = [\lambda(z-1) - \mu_k]G_{X_{(k)}}(t,z) + \mu_k G_{X_{(k-1)}}(t,z). \tag{44}$$

B. Load analysis of $\mathbf{Y}(s)$

Let $Y_{(k)}(s)$ denote the total number of particles present, immediately after the sth Poissonian event ($s = 1,2,\ldots$), in the first k waiting zones. The random variable $Y_{(k)}(s)$ is the sum of the first k coordinates of the random vector $\mathbf{Y}(s)$, i.e., $Y_{(k)}(s) = Y_1(s) + \cdots + Y_k(s)$ ($k = 1,\ldots,n$).

Observe the system at two consecutive s and $s' = s + 1$ Poissonian events, and use the shorthand notation $Y_{(k)} = Y_{(k)}(s)$ and $Y'_{(k)} = Y_{(k)}(s')$ ($k = 1,\ldots,n$). Equation (5) implies that the stochastic connection between the random sums $Y_{(k)}$ and $Y'_{(k)}$, characterizing the law of motion of the stochastic process $[Y_{(k)}(s)]_{s=1}^{\infty}$, is given by

$$Y'_{(k)} = \begin{cases} Y_{(k)} & \text{w.p. } \frac{\mu-\mu_k}{\lambda+\mu}, \\ Y_{(k)} + 1 & \text{w.p. } \frac{\lambda}{\lambda+\mu}, \\ Y_{(k-1)} & \text{w.p. } \frac{\mu_k}{\lambda+\mu}. \end{cases} \tag{45}$$

Equation (45) follows from considering the events that result in a change $Y_{(k)} \mapsto Y'_{(k)} \neq Y_{(k)}$. There are exactly two such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first waiting zone, occurring with probability $\lambda/(\lambda + \mu)$, in which case $Y_{(k)} \mapsto Y'_{(k)} = Y_{(k)} + 1$; ($k$) opening of the kth gate, occurring with probability $\mu_k/(\lambda + \mu)$, in which case $Y_{(k)} \mapsto Y'_{(k)} = Y_{(k-1)}$. The first line on the right-hand side of Eq. (45) represents the scenario in which a gate other than the kth gate opens, which occurs with the complementary probability $(\mu - \mu_k)/(\lambda + \mu)$.

Let

$$G_{Y_{(k)}}(s,z) = \mathrm{E}[z^{Y_{(k)}(s)}] \tag{46}$$

($|z| \leqslant 1$) denote the PGF of the random sum $Y_{(k)}(s)$. Setting $z_1 = \cdots = z_k = z$ and $z_{k+1} = \cdots = z_n = 1$, and noting that, by definition, $G_{Y_{(k)}}(s,z) = G_{\mathbf{Y}}(s,z,\ldots,z,1,\ldots,1)$, Eq. (25) yields

$$G_{Y_{(k)}}(s',z) - G_{Y_{(k)}}(s,z) = \begin{cases} \frac{\lambda(z-1)-\mu}{\lambda+\mu} G_{Y_{(k)}}(s,z) \\ + \\ \frac{\mu_1}{\lambda+\mu} G_{Y_{(k)}}(s,z) \\ + \cdots + \\ \frac{\mu_{k-1}}{\lambda+\mu} G_{Y_{(k)}}(s,z) \\ + \\ \frac{\mu_k}{\lambda+\mu} G_{Y_{(k-1)}}(s,z) \\ + \\ \frac{\mu_{k+1}}{\lambda+\mu} G_{Y_{(k)}}(s,z) \\ + \cdots + \\ \frac{\mu_{n-1}}{\lambda+\mu} G_{Y_{(k)}}(s,z) \\ + \\ \frac{\mu_n}{\lambda+\mu} G_{Y_{(k)}}(s,z). \end{cases} \tag{47}$$

Equation (47), in turn, implies that the PGF dynamics of the random sum $Y_{(k)}(s)$ is given by

$$G_{Y_{(k)}}(s',z) - G_{Y_{(k)}}(s,z) = \frac{\lambda(z-1) - \mu_k}{\lambda+\mu} G_{Y_{(k)}}(s,z) + \frac{\mu_k}{\lambda+\mu} G_{Y_{(k-1)}}(s,z). \tag{48}$$

C. Steady state

Consider now the ASIP model in *steady state*. In steady state, the stochastic processes $[X_{(k)}(t)]_{t\geqslant 0}$ and $[Y_{(k)}(s)]_{s=1}^{\infty}$ are *stationary*, and hence their respective PGFs are time homogeneous: $G_{X_{(k)}}(t,z) \equiv G_{X_{(k)}}(z)$ ($t \geqslant 0$) and $G_{Y_{(k)}}(s,z) \equiv G_{Y_{(k)}}(z)$ ($s = 1,2,\ldots$). Substituting the time-homogeneous PGFs $G_{X_{(k)}}(t,z) \equiv G_{X_{(k)}}(z)$ and $G_{Y_{(k)}}(s,z) \equiv G_{Y_{(k)}}(z)$, respectively, into Eqs. (44) and (48) yields the common equation

$$G_k(z) = \frac{\mu_k}{\mu_k + \lambda(1-z)} G_{k-1}(z) \tag{49}$$

($k = 1,\ldots,n$). Namely, both the PGFs $G_{X_{(k)}}(z)$ and $G_{Y_{(k)}}(z)$ are governed by Eq. (49).

Note that $X_{(1)}(t) = X_1(t)$ and $Y_{(1)}(s) = Y_1(s)$ and, hence, the PGF $G_1(z)$ is given by Eq. (29). Using the initial condition

$G_1(z)$ and iterating Eq. (49), we obtain that

$$\mathrm{E}[z^{X_{(k)}(t)}] = \mathrm{E}[z^{Y_{(k)}(s)}] = \frac{\mu_1}{\mu_1+\lambda(1-z)}\cdots\frac{\mu_k}{\mu_k+\lambda(1-z)} = \frac{p_1}{1-(1-p_1)z}\cdots\frac{p_k}{1-(1-p_k)z}, \tag{50}$$

where $p_k = \mu_k/(\mu_k+\lambda)$ $(k=1,\ldots,n)$. As the initial condition implies, in the case $k=1$, Eq. (50) coincides with Eq. (29). Interestingly, for $k>1$, Eq. (50) attains a *product-form representation*. This product form implies that both $X_{(k)}(t)$ and $Y_{(k)}(t)$ are characterized by the following *stochastic decomposition*: The random variables $X_{(k)}(t)$ and $Y_{(k)}(t)$ are equal, in law, to the total number of particles in k *independent*and *single gated* ASIP systems with respective parameters $(\lambda,\mu_1),\ldots,(\lambda,\mu_k)$.

The PGF $G(z) = p/[1-(1-p)z]$ characterizes a *geometric law* on the non-negative integers, which, in turn, has mean $\frac{1-p}{p}$ and variance $\frac{1-p}{p^2}$. Combining this fact together with the aforementioned stochastic representation, we obtain that the mean and variance of the random variables $X_{(k)}(t)$ and $Y_{(k)}(t)$ are given, respectively, by

$$\mathrm{E}[X_{(k)}(t)] = \mathrm{E}[Y_{(k)}(s)] = \lambda\left(\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_k}\right) \tag{51}$$

and

$$\mathrm{Var}[X_{(k)}(t)] = \mathrm{Var}[Y_{(k)}(s)] = \lambda\left(\frac{\mu_1+\lambda}{\mu_1^2}+\cdots+\frac{\mu_k+\lambda}{\mu_k^2}\right). \tag{52}$$

We emphasize that Eq. (50) implies that the distribution of $X_{(n)}(t)$ and $Y_{(n)}(s)$ is *independent of the order of the gates*. Consequently, permuting the gates (each gate carrying its own opening rate with it) has no effect on the distribution of the ASIP load. Thus, from a load perspective, the ASIP model is invariant with respect to gate permutations.

VII. LOAD OPTIMIZATION

To design an efficient ASIP system, one would like to minimize the system's load, i.e., to minimize the number of particles "in process" termed "work-in-process" (WIP in production models [45]). In this section, we explore the optimization of the ASIP's load. In what follows, we consider as given the exogenous inflow rate λ, and optimize the endogenous service rates $\{\mu_1,\mu_2,\ldots,\mu_n\}$.

A. Optimality

We begin with the *combinatorial optimization* of the ASIP model. Namely, given a collection of n gates, each with its own service rate, we seek an ordering of gates that renders a target-functional optimal. As explained in the previous section, the distribution of the ASIP's load is invariant with respect to gate permutations. Hence, for any target functional based on the ASIP's load distribution, optimization is trivial: all gate permutations yield the same target-functional value.

Let us turn now to examine *constrained optimization* of the ASIP model. To that end, we consider four optimization problems in which we seek to minimize a target functional based on the ASIP's load distribution, subject to a given constraint.

(i) *Minimization of the load mean subject to a given cumulative service rate*. Assume that the cumulative service rate μ is fixed and constant. Here, we seek an optimal allocation of the cumulative service rate μ to the different gates, the goal being a minimal load mean. Recalling Eq. (17), which asserts that the load mean is given by the product $\lambda\mathrm{E}[T]$, we note that the minimization of the load mean is equivalent to the minimizing of the traversal time. By applying Eq. (51), we obtain the constrained optimization problem

$$\begin{aligned}\min\quad & \left\{\lambda\left(\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_n}\right)\right\}\\ \text{s.t.}\quad & \mu_1+\cdots+\mu_n=\mu,\end{aligned} \tag{53}$$

where s.t. stand for "subject to".

(ii) *Minimization of the load variance subject to a given cumulative service rate*. Assume that the cumulative service rate μ is fixed and constant. Here, we seek an optimal allocation of the cumulative service rate μ to the different gates, the goal being a minimal load variance. By applying Eq. (52), we obtain the constrained optimization problem

$$\begin{aligned}\min\quad & \left\{\lambda\left(\frac{\mu_1+\lambda}{\mu_1^2}+\cdots+\frac{\mu_n+\lambda}{\mu_n^2}\right)\right\}\\ \text{s.t.}\quad & \mu_1+\cdots+\mu_n=\mu.\end{aligned} \tag{54}$$

(iii) *Minimization of the load variance subject to a given load mean*. Assume that the load mean is predetermined to equal the value v (alternatively, assume that the traversal time is predetermined to equal the value v/λ). Here, we seek the optimal service rates that render the load variance minimal. By applying Eqs. (51) and (52), we obtain the constrained optimization problem

$$\begin{aligned}\min\quad & \left\{\lambda\left(\frac{\mu_1+\lambda}{\mu_1^2}+\cdots+\frac{\mu_n+\lambda}{\mu_n^2}\right)\right\}\\ \text{s.t.}\quad & \lambda\left(\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_n}\right)=v.\end{aligned} \tag{55}$$

This optimization problem is analogous to the Markowitz optimization of financial portfolios in which one seeks to minimize the portfolio variance, subject to a predetermined portfolio mean [46].

(iv) *Maximization of the zero-load probability subject to a given cumulative service rate*. Assume that the cumulative service rate is fixed and constant. Here, we seek an optimal allocation of the cumulative service rate to the different gates, the goal being a maximal zero-load probability $\Pr[X_{(n)}(t)=0]$. This zero-load probability is attained by setting $z=0$ into the PGF of the load $X_{(n)}(t)$. By setting $z=0$ into the right-hand side of Eq. (50), we obtain the constrained optimization problem

$$\begin{aligned}\max\quad & \left\{\frac{\mu_1}{\mu_1+\lambda}\cdots\frac{\mu_n}{\mu_n+\lambda}\right\}\\ \text{s.t.}\quad & \mu_1+\cdots+\mu_n=\mu.\end{aligned} \tag{56}$$

Note that the constrained optimization problem (56) is equivalent to the constrained optimization problem

$$\begin{aligned} \min \quad & \left\{ \ln\left(1+\frac{\lambda}{\mu_1}\right) + \cdots + \ln\left(1+\frac{\lambda}{\mu_n}\right) \right\} \\ \text{s.t.} \quad & \\ & \mu_1 + \cdots + \mu_n = \mu. \end{aligned} \tag{57}$$

The four aforementioned optimization problems admit the general form

$$\begin{aligned} \min \quad & \{f(x_1) + \cdots + f(x_n)\} \\ \text{s.t.} \quad & \\ & x_1 + \cdots + x_n = c, \end{aligned} \tag{58}$$

where $f(x)$ is a convex function and the variables are positive valued: $x_1, \ldots, x_n > 0$. Indeed, (i) in the first problem, $x_k = \mu_k$, $c = \mu$, and $f(x) = \lambda/x$; (ii) in the second problem, $x_k = \mu_k$, $c = \mu$, and $f(x) = (\lambda/x) + (\lambda/x)^2$; (iii) in the third problem, $x_k = 1/\mu_k$, $c = v/\lambda$, and $f(x) = (\lambda x) + (\lambda x)^2$; and (iv) in the fourth problem, $x_k = \mu_k$, $c = \mu$, and $f(x) = \ln(1 + \lambda/x)$. The Lagrange function corresponding to the optimization problem of Eq. (58) is given by

$$L(x_1, \ldots, x_n; \theta) = \left(\sum_{k=1}^{n} f(x_k) \right) + \theta \left(c - \sum_{k=1}^{n} x_k \right). \tag{59}$$

Differentiating the Lagrange function with respect to the variable x_k and equating the partial derivative to zero yields the equation

$$f'(x_k) = \theta. \tag{60}$$

Now, since the function $f(x)$ is convex $[f''(x) > 0]$, its derivative $f'(x)$ is monotone increasing. This implies that Eq. (60) admits a unique solution, which, in turn, implies that the unique critical point of the Lagrange function satisfies $x_1 = \cdots = x_n$. Since the target function $\sum_{k=1}^{n} f(x_k)$ is convex, and the constraint $\sum_{k=1}^{n} x_k = c$ is linear, we conclude that [47] the global minimum of the optimization problem (58) is given by $x_1 = \cdots = x_n = c/n$.

Thus, the solution to all four aforementioned optimization problems turns out to be a *homogeneous ASIP system*, with service rates $\mu_1 = \mu_2 = \cdots = \mu_n$. This optimization conclusion highlights the importance of homogeneous ASIP systems within the class of general ASIP systems. The optimality of the homogeneous solution is illustrated graphically in Fig. 5.

B. Deviations from optimality and bottlenecks

Having concluded that homogeneous ASIP systems are optimal, we turn to discuss deviations from optimality. Of particular interest is the sensitivity of the target function to small changes in the optimal service rates vector. To this end, we find it useful to borrow the "bottleneck" concept from the ASEP nomenclature [48]. Bottlenecks are sites where the hopping rate of particles is reduced compared to the rest of the system. In the ASEP, the main effect of bottlenecks is to decrease the current (or flow) through the system [49]. In the ASIP, the steady-state flow of particles is always given by λ and, hence, is independent of the service rates $\{\mu_1, \mu_2, \ldots, \mu_n\}$. Interestingly, bottlenecks are nevertheless

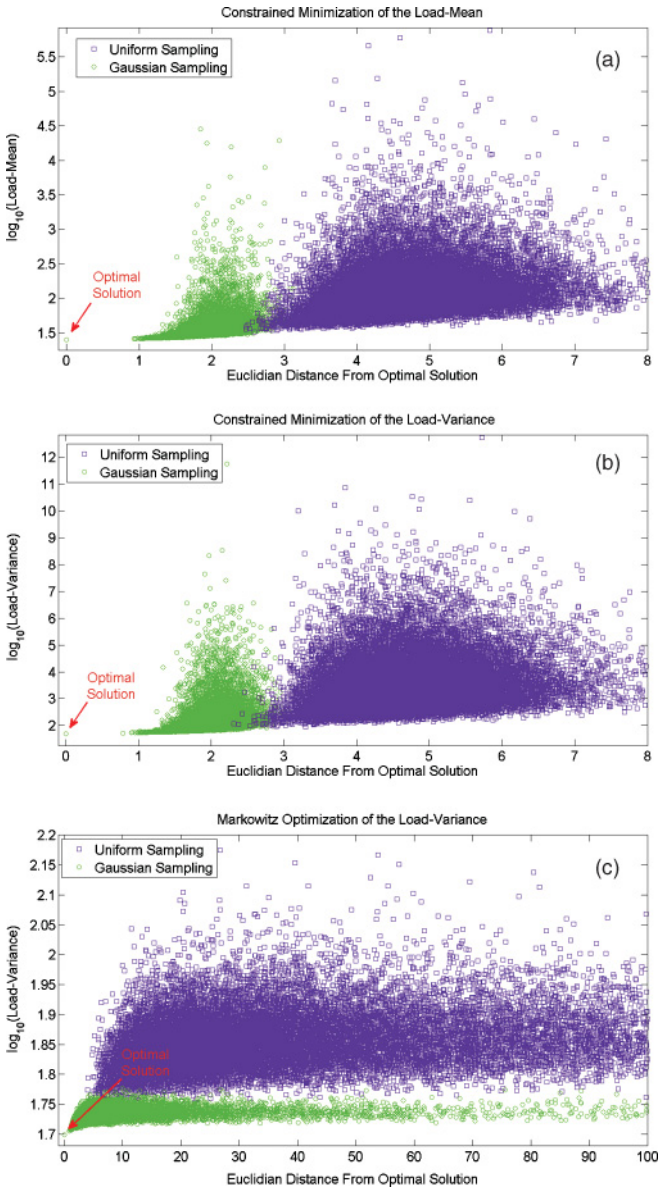

FIG. 5. (Color online) Optimality of homogeneous ASIP systems. Panels (a)–(c) are associated with the optimization problems presented in Eqs. (53), (54), and (56), respectively. In all three panels, results are shown for ASIP systems with 25 gates ($n = 25$) and an inflow rate of $\lambda = 1$. The constraint parameters μ in Eqs. (53) and (54) and v in Eq. (56) are taken to equal 25 ($\mu = v = 25$). The optimal solution under these conditions is identical for all three problems and is given by $\mu_1 = \mu_2 = \cdots = \mu_{25} = 1$. The value of the target function, evaluated at randomly drawn rate vectors $(\mu_1, \ldots, \mu_{25})$, is plotted vs the Euclidean distance of these vectors from the optimal rate vector $(1, 1, \ldots, 1)$. The optimality of the latter is clearly visible. In each panel, rate vectors are randomly drawn 25 000 times in the two following methods: Gaussian sampling and uniform sampling. In the Gaussian sampling, Gaussian noise is added to the optimal rate vector. This vector is then normalized to form a rate vector that complies with the problem constraints. In the uniform sampling, the interval $[0, 25]$ is dissected into 25 segments by randomly drawing 24 numbers from a uniform distribution over that interval. The lengths of these segments are then taken to represent the rate vector [in the case of panels (a) and (b)] or the inverse rate vector $(\frac{1}{\mu_1}, \ldots, \frac{1}{\mu_{25}})$ in the case of panel (c).

useful in understanding deviations from optimality since both the load mean and load variance of an ASIP system are sensitive to their existence.

When it comes to the sensitivity of the target function to perturbations around the optimal solution, Fig. 5(c) shows a strikingly different behavior compared to Figs. 5(a) and 5(b). While in Figs. 5(a) and 5(b), small deviations from the optimal rates vector may change the target function by orders of magnitude, this is hardly the case in the Fig. 5(c).

This phenomenon can easily be understood by noting that in the first two optimization problems discussed above, the given cumulative service rate constraint does not impose a lower bound on the service rates in the system. This constraint can therefore be satisfied even in the presence of a site, the service rate of which is infinitesimally small. A single bottleneck (or defect) within an otherwise (almost) homogeneous ASIP system will result in nothing but a slight deviation from the optimal solution. However, since the load mean and load variance are highly sensitive to the existence of bottlenecks, the impact on the target function will be tremendous.

The situation is considerably different when the given cumulative service rate constraint is replaced by a given load-mean constraint as is done in the third optimization problem above. The latter imposes a lower bound on the service rates in the system. Moreover, in order to satisfy the constraint, the existence of a bottleneck forces the allocation of extremely high service rates to many other sites. And so, the impact of bottlenecks on the target function is both limited and, when substantial, accompanied by a discernible deviation from the optimal solution.

VIII. CONCLUSIONS

We introduced and analyzed the asymmetric inclusion process (ASIP), a model for a unidirectionally driven lattice gas of particles subject to inclusion interactions. The ASIP represents a bosonic counterpart of the fermionic asymmetric exclusion process (ASEP), a tandem array of Markovian queueing systems with (unlimited) batch service and a tandem array of growth-collapse processes. The ASIP, counterwise to its simple description, is highly complex and intractable. Nonetheless, we succeeded in penetrating the ASIP's intractability by obtaining the following results: (i) explicit evolution equations for the mean and PGF; (ii) explicit solution of the mean in steady state; (iii) explicit equations for the PGF in steady state; (iv) explicit solution of the steady-state PGF for small systems ($n = 1,2,3$), and an iterative scheme for the computation of the steady-state PGF for systems of arbitrary size; (v) explicit solution of the mean, variance, and PGF of the load in steady state; and (vi) explicit solutions of various load-optimization problems, rendering homogeneous ASIP models optimal. Our paper joins a recent and novel set of research papers bridging the seemingly distinct fields of statistical physics and queueing theory.

ACKNOWLEDGMENT

S.R. acknowledges support from the Converging Technologies program of the Israeli Council for higher education.

APPENDIX

We solve (27) for the ASIP with three gates $n = 3$. In this case, Eq. (27) reduces to

$$[\lambda(1-z_1)+\mu]G(z_1,z_2,z_3) = \begin{cases} \mu_1 G(z_2,z_2,z_3) \\ + \\ \mu_2 G(z_1,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_2,1). \end{cases} \tag{A1}$$

Now, following the scheme's basic step, we iteratively apply Eq. (A1) to the daughters $G(z_2,z_2,z_3)$, $G(z_1,z_3,z_3)$, and $G(z_1,z_2,1)$.

From the embedding property (see Sec. IV C), the daughter $G(z_1,z_2,1)$ is equal to $G(z_1,z_2)$ and is hence known and given by Eq. (39). For the daughter $G(z_1,z_3,z_3)$, the basic step yields

$$[\lambda(1-z_1)+\mu]G(z_1,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_2 G(z_1,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_3,1) \end{cases} \tag{A2}$$

from which we obtain

$$[\lambda(1-z_1)+\mu_1+\mu_3]G(z_1,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_3,1). \end{cases} \tag{A3}$$

Again, the daughter $G(z_1,z_3,1)$ is known and given by Eq. (39). For the daughter $G(z_3,z_3,z_3)$, the basic step yields

$$[\lambda(1-z_3)+\mu]G(z_3,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_2 G(z_3,z_3,z_3) \\ + \\ \mu_3 G(z_3,z_3,1) \end{cases} \tag{A4}$$

from which we obtain

$$[\lambda(1-z_3)+\mu_3]G(z_3,z_3,z_3) = \mu_3 G(z_3,z_3,1). \tag{A5}$$

We conclude that

$$G(z_1,z_3,z_3) = \begin{cases} \frac{\mu_1\mu_3 G(z_3,z_3,1)}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_3)+\mu_3]} \\ + \\ \frac{\mu_3 G(z_1,z_3,1)}{\lambda(1-z_1)+\mu_1+\mu_3}. \end{cases} \tag{A6}$$

We now return to the daughter $G(z_2,z_2,z_3)$; applying the basic step yields

$$[\lambda(1-z_2)+\mu]G(z_2,z_2,z_3) = \begin{cases} \mu_1 G(z_2,z_2,z_3) \\ + \\ \mu_2 G(z_2,z_3,z_3) \\ + \\ \mu_3 G(z_2,z_2,1) \end{cases} \tag{A7}$$

from which we obtain

$$[\lambda(1-z_2)+\mu_2+\mu_3]G(z_2,z_2,z_3) = \begin{cases} \mu_2 G(z_2,z_3,z_3) \\ + \\ \mu_3 G(z_2,z_2,1). \end{cases} \tag{A8}$$

Here, both the daughter $G(z_2,z_2,1)$ and the daughter $G(z_2,z_3,z_3)$ are known and given by Eqs. (38) and (A6), respectively. We conclude that

$$G(z_2,z_2,z_3) = \begin{cases} \frac{\mu_1\mu_2\mu_3 G(z_3,z_3,1)}{[\lambda(1-z_2)+\mu_1+\mu_3][\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_3)+\mu_3]} \\ + \\ \frac{\mu_2\mu_3 G(z_2,z_3,1)}{[\lambda(1-z_2)+\mu_1+\mu_3][\lambda(1-z_2)+\mu_2+\mu_3]} \\ + \\ \frac{\mu_3 G(z_2,z_2,1)}{\lambda(1-z_2)+\mu_2+\mu_3}. \end{cases} \tag{A9}$$

By substituting the expressions for $G(z_2,z_2,z_3)$ and $G(z_1,z_3,z_3)$ into Eq. (A1), we obtain

$$[\lambda(1-z_1)+\mu]G(z_1,z_2,z_3) = \begin{cases} \frac{\mu_1^2\mu_2\mu_3 G(z_3,z_3,1)}{[\lambda(1-z_2)+\mu_1+\mu_3][\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_3)+\mu_3]} \\ + \\ \frac{\mu_1\mu_2\mu_3 G(z_2,z_3,1)}{[\lambda(1-z_2)+\mu_1+\mu_3][\lambda(1-z_2)+\mu_2+\mu_3]} \\ + \\ \frac{\mu_1\mu_3 G(z_2,z_2,1)}{\lambda(1-z_2)+\mu_2+\mu_3} \\ + \\ \frac{\mu_1\mu_2\mu_3 G(z_3,z_3,1)}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_3)+\mu_3]} \\ + \\ \frac{\mu_2\mu_3 G(z_1,z_3,1)}{\lambda(1-z_1)+\mu_1+\mu_3} \\ + \\ \mu_3 G(z_1,z_2,1). \end{cases} \tag{A10}$$

By substituting the expressions for $G(z_3,z_3,1)$, $G(z_2,z_3,1)$, $G(z_2,z_2,1)$, $G(z_1,z_3,1)$, and $G(z_1,z_2,1)$ into Eq. (A10), we obtain the final expression for $G(z_1,z_2,z_3)$ given in Eq. (40).

[1] V. Y. Chernyak, M. Chertkov, D. A. Goldberg, and K. Turitsyn, J. Stat. Phys. **140**, 819 (2010).
[2] N. Merhav and Y. Kafri, J. Stat. Mech. (2010) P02011.
[3] A. Arazi, E. Ben-Jacob, and U. Yechiali, Phys. A (Amsterdam) **332**, 585 (2004).
[4] A. Arazi and U. Yechiali (unpublished).
[5] B. Derrida, Phys. Rep. **301**, 65 (1998).
[6] O. Golinelli and K. Mallick, J. Phys. A: Math. Gen. **39**, 12679 (2006).
[7] M. F. Neuts, Oper. Res. **13**, 815 (1965).
[8] H. Kaspi, O. Kella, and D. Perry, Queueing Syst. Theory Appl. **24**, 37 (1997).
[9] I. Eliazar and J. Klafter, Phys. A (Amsterdam) **334**, 1 (2004).
[10] O. Boxma, D. Perry, W. Stadje, and S. Zacks, Adv. Appl. Probab. **38**, 221 (2006).
[11] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, Biopolymers **6**, 1 (1968).
[12] K. Heckmann, *Biomembranes 3*, edited by F. Kreuzer and J. F. G. Slegers (Plenum, New York, 1972).
[13] D. G. Levitt, Phys. Rev. A **8**, 3050 (1973).
[14] P. M. Richards, Phys. Rev. B **16**, 1393 (1977).
[15] B. Widom, J. L. Viovy, and A. D. Defontaines, J. Phys. I **1**, 1759 (1991).
[16] *Traffic and Granular Flow '97*, edited by M. Schreckenberg and D. E. Wolf (Springer, New York, 1998).
[17] L. B. Shaw, R. K. Zia, and K. H. Lee, Phys. Rev. E **68**, 021910 (2003).
[18] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller, PLoS Comput. Biol. **7**(9), e1002127 (2011).
[19] T. Halpin-Healy and Y. C. Zhang, Phys. Rep. **254**, 215 (1995).
[20] J. Krug, Adv. Phys. **46**, 139 (1997).
[21] R. Bundschuh, Phys. Rev. E **65**, 031911 (2002).
[22] S. Klumpp and R. Lipowsky, J. Stat. Phys. **113**, 233 (2003).
[23] S. F. Burlatsky, G. S. Oshanin, A. V. Mogutov, and M. Moreau, Phys. Lett. A **166**, 230 (1992).
[24] S. F. Burlatsky, G. Oshanin, M. Moreau, and W. P. Reinhardt, Phys. Rev. E **54**, 3165 (1996).
[25] O. Bénichou, A. M. Cazabat, J. De Coninck, M. Moreau, and G. Oshanin, Phys. Rev. Lett. **84**, 511 (2000).
[26] C. M. Monasterio and Gleb Oshanin, Soft Matter **7**, 993 (2011).
[27] R. A. Blythe and M. R. Evans, J. Phys. A: Math. Theor. **40**, R333 (2007).
[28] J. R. Jackson, Management Sci. **10**, 131 (1963).
[29] J. R. Jackson, Oper. Res. **5**, 518 (1957).
[30] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks* (Springer, Berlin, 2001).
[31] K. Nishinari and D. Takahashi, J. Phys. A: Math. Gen. **31**, 5439 (1998).
[32] F. Spitzer, Adv. Math. **5**, 246 (1970).
[33] E. Levine, D. Mukamel, and G. M. Schutz, J. Stat. Phys. **120**, 5/6 (2005).
[34] M. R. Evans and T. Hanney, J. Phys. A: Math. Gen. **38**, R195 (2005).
[35] O. Kella, J. Appl. Probab. **46**, 363 (2009).
[36] B. J. Martin, Queueing Syst. Theory Appl. **62**, 411 (2009).
[37] P. Bak, *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York, 1996).
[38] M. G. Rozman, M. Urbach, J. Klafter, and F. J. Elmer, J. Phys. Chem. B **102**, 7924 (1998).
[39] J. M. Carlson, J. S. Langer, and B. E. Shaw, Rev. Mod. Phys. **66**, 657 (1994).
[40] I. Eliazar and J. Klafter, J. Stat. Phys. **118**, 177 (2005).
[41] I. Eliazar and J. Klafter, Phys. A (Amsterdam) **367**, 106 (2006).
[42] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., Vol. 2 (Wiley, New York, 1991).

[43] R. W. Wolff, *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).
[44] I. Eliazar and J. Klafter, Phys. A (Amsterdam) **388**, 1755 (2009).
[45] R. G. Bitran and D. Tirupati, Oper. Res. **37**, 547 (1989).
[46] H. M. Markowitz, J. Finance **7**, 77 (1952).
[47] D. Bertsekas, *Convex Analysis and Optimization* (Athena Scientific, Belmont, MA, 2003).
[48] P. Greulich and A. Schadschneider, Phys. A (Amsterdam) **387**, 1972 (2008).
[49] A. B. Kolomeisky, J. Phys. A: Math. Gen. **31**, 1153 (1998).