Queueing Syst (2008) 59: 135–156
DOI 10.1007/s11134-008-9079-4

The asymptotic variance rate of the output process of finite capacity birth-death queues

Yoni Nazarathy · Gideon Weiss

Received: 16 January 2008 / Revised: 8 July 2008 / Published online: 6 August 2008

Abstract We analyze the output process of finite capacity birth-death Markovian queues. We develop a formula for the asymptotic variance rate of the form $\lambda^* + \sum v_i$ where λ^* is the rate of outputs and v_i are functions of the birth and death rates. We show that if the birth rates are non-increasing and the death rates are non-decreasing (as is common in many queueing systems) then the values of v_i are strictly negative and thus the limiting index of dispersion of counts of the output process is less than unity.

In the M/M/1/K case, our formula evaluates to a closed form expression that shows the following phenomenon: When the system is balanced, i.e. the arrival and service rates are equal, $\frac{\sum v_i}{\lambda^*}$ is minimal. The situation is similar for the M/M/c/K queue, the Erlang loss system and some PH/PH/1/K queues: In all these systems there is a pronounced decrease in the asymptotic variance rate when the system parameters are balanced.

Keywords Queueing theory · Loss systems · M/M/1/K · MAP · Asymptotic variance rate · BRAVO

Mathematics Subject Classification (2000) 60J27 · 60K25

1 Introduction

Let $\mathcal{Q} = \{Q(t), t \geq 0\}$ be the number of jobs in a queueing system and assume that it is an irreducible, stationary continuous time Markov chain (CTMC) with a birth-death structure on the finite state space $\{0, \ldots, K\}$. Let $\mathcal{D} = \{D(t), t \geq 0\}$ be the

Research supported in part by Israel Science Foundation Grant 249/02 and 454/05 and by European Network of Excellence Euro-NGI.

Y. Nazarathy (✉) · G. Weiss
Department of Statistics, The University of Haifa, Mount Carmel 31905, Israel
e-mail: yonin@stat.haifa.ac.il

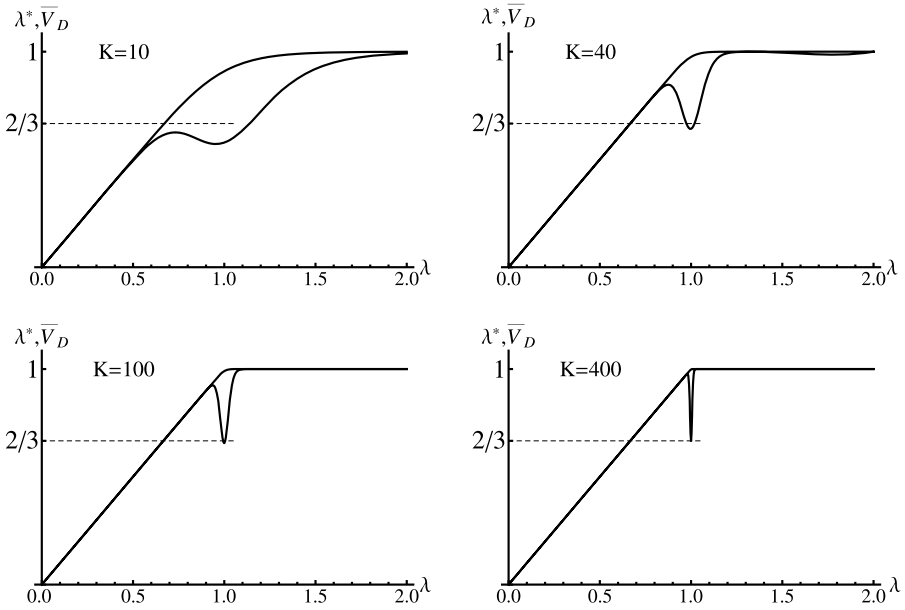

Fig. 1 M/M/1/K: λ^* (*top curve*) and $\overline{V}_{\mathcal{D}}$ (*bottom curve*) as a function of λ when $\mu = 1$ for various buffer sizes

output process associated with the queue: $D(0) = 0$ and $\mathcal{D}$ increases by 1 when $\mathcal{Q}$ decreases. It can be shown that the expectation and the variance functions of $\mathcal{D}$ are $O(t)$ (cf. formulas (9), (10) and accompanying discussion and references) and may thus be described by the *flow rate*, λ^* and *asymptotic variance rate*, $\overline{V}_{\mathcal{D}}$:

$$\mathbb{E}\,[D(t)] = \lambda^* t \tag{1}$$

$$\mathrm{Var}(D(t)) = \overline{V}_{\mathcal{D}} t + o(t) \tag{2}$$

Evaluation of $\overline{V}_{\mathcal{D}}$ is important in manufacturing type settings. When the system operates for a long duration, T, the variance of the number of items produced is approximately $\overline{V}_{\mathcal{D}} T$. Several studies have investigated computational procedures that evaluate this quantity for the output of a series of queues (cf. [14, 16, 17, 21, 29, 30]). In this paper we concentrate on the seemingly simpler case of a one-pass single class queueing system with losses. In general, output processes of one-pass single class systems and their second moments have been studied extensively (cf. the surveys [9, 12, 27]). For finite state space loss systems, the overflow process has received a considerable amount of attention (cf. [3, 4, 7, 24, 26, 31, 36]). Fewer papers have considered the output process of loss systems (cf. [2, 10, 23]) and to the best of our knowledge none have analyzed the asymptotic variance rate of the outputs.

Figures 1 and 2 display $\overline{V}_{\mathcal{D}}$ for different parameter values of the M/M/1/K queue with arrival rate λ and service rate μ. The plots may be partially understood as follows: For $\lambda \ll \mu$ the finite queue is hardly ever full and it behaves almost like an M/M/1 queue. In the M/M/1 queue, reversibility arguments imply that $\mathcal{D}$ is a Poisson process (cf. [19]), and thus for M/M/1/K we expect $\overline{V}_{\mathcal{D}} \approx \lambda^* \approx \lambda$ when $\lambda \ll \mu$. For

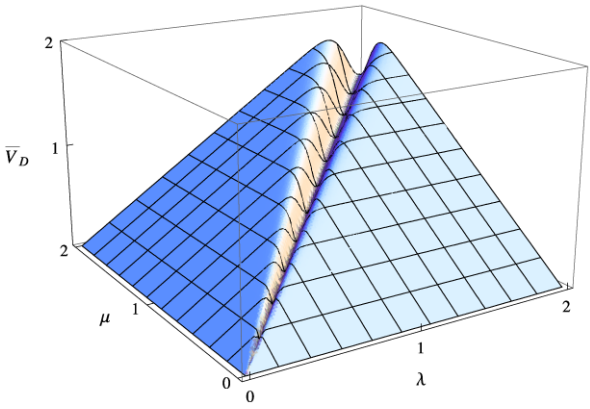

Fig. 2 M/M/1/40: $\overline{V}_{\mathcal{D}}$ as a function of λ and μ

$\lambda \gg \mu$ the queue is almost always full and thus the outputs are similar to a Poisson process with rate μ so we expect $\overline{V}_{\mathcal{D}} \approx \lambda^* \approx \mu$ when $\lambda \gg \mu$. The behavior of the plots of $\overline{V}_{\mathcal{D}}$ when $\lambda \approx \mu$ is not easily explained: There is a pronounced decrease to a value of approximately $\frac{2}{3}\lambda$. To the best of our knowledge, this phenomenon has not been documented previously. We loosely refer to this as the **BRAVO** effect which stands for: **B**alancing **R**educes **A**symptotic **V**ariance of **O**utputs. Our results show that BRAVO occurs in a variety of finite capacity queueing models with losses.

Still focusing on the M/M/1/K as an example, notice that while it can be shown that the process which is the sum of the outputs and the overflows is Poisson, $\mathcal{D}$ by itself is not Poisson. One may attempt to evaluate the asymptotic variance rate by treating $\mathcal{D}$ as a renewal process. In this case $\overline{V}_{\mathcal{D}} = c^2\lambda^*$ where $c^2 = \frac{\sigma^2}{m^2}$ denotes the squared coefficient of variation (SCV) of the stationary inter-output time having expectation m and variance σ^2 (cf. [1], p. 161). Variations of this method have been used to approximate inter-node flows in queueing networks (cf. [32, 33] and references therein). But it is known that the output process of most finite buffer queueing systems is not a renewal process (cf. [12], Sect. VII) and thus there is no theoretical justification for approximating $\overline{V}_{\mathcal{D}}$ using a renewal process. In fact, this type of approximation may yield completely incorrect results when the service rate and arrival rate are similar. For example, in the M/M/1/K queue case, the renewal approximation yields $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*} = 1$ for $\lambda = \mu$, while the actual value is nearly $\frac{2}{3}$.

The probability law of $\mathcal{D}$ has been thoroughly researched. It is a Markov Renewal Process and also a Markovian Arrival Process (MAP) (cf. [1, 11]). It is possible numerically to compute $\overline{V}_{\mathcal{D}}$, and even $\mathrm{Var}(D(t))$ for any t, using well established matrix analytic results (see formulas (10) and (11) and references [22, 23]). Thus, discovery of BRAVO did not require any new machinery beyond formula (10).

Our results are as follows: Part (i) of our main theorem (Theorem 3.1) is the formula $\overline{V}_{\mathcal{D}} = \lambda^* + \sum_{i=0}^{K-1} v_i$, where v_i, $i = 0, \ldots, K-1$ are expressions based on the birth and death rates. When applied to the M/M/1/K queue this formula yields a simple closed form expression. Part (ii) shows that when the birth rates are non-increasing and the death rates are non-decreasing (as is the case in many queueing systems), $v_i < 0$ for $i = 0, \ldots, K-1$ and hence, $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*}$, the *limiting index of dispersion of counts* (cf. [8]) is less than unity. For M/M/1/K queue we also derive additional

results: an expression for the asymptotic correlation between the output and overflow processes and an expression for the y-intercept of the linear asymptote of $\mathrm{Var}(D(t))$ for the balanced case.

The proof of Part (i) of our main theorem relies on a complementary result (Proposition 3.2) which relates to a class of MAPs that count every transition of a CTMC. We show that such MAPs have an associated Markov Modulated Poisson Process (MMPP) which has the same expectation and variance functions as the original MAP. This result may be of independent interest.

The rest of the paper is organized as follows: In Sect. 2 we present some notation and fundamental results that are used throughout. In Sect. 3 we state and prove our main theorem. In Sect. 4 we analyze the M/M/1/K queue. In Sect. 5 we show that the BRAVO effect also appears in M/M/c/K systems (including the Erlang loss system), and in some PH/PH/1/K queues which we analyze numerically. Some remaining open questions are also discussed.

2 Preliminaries

We now introduce further notation and preliminary results that are needed in the paper.

Birth-death CTMCs We assume throughout that $\mathcal{Q}$ is a finite state space stationary birth-death process with generator matrix:

$$\mathbf{\Lambda} = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & 0 \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{K-1} & -(\mu_{K-1}+\lambda_{K-1}) & \lambda_{K-1} \\ 0 & & & \mu_K & -\mu_K \end{pmatrix} \tag{3}$$

The birth rates are $\lambda_0, \ldots, \lambda_{K-1} > 0$ and the death rates are $\mu_1, \ldots, \mu_K > 0$. The stationary probability distribution $\boldsymbol{\pi} = \{\pi_i, i = 0, \ldots, K\}$ is the solution of the equations: $\boldsymbol{\pi}\mathbf{\Lambda} = \mathbf{0}$, $\boldsymbol{\pi}\mathbf{1} = 1$, where we take $\boldsymbol{\pi}$ to be a row vector, $\mathbf{0}$ to be a row vector of 0s and $\mathbf{1}$ to be column vectors of 1s. It is well known that the stationary distribution is:

$$\pi_i = \frac{\lambda_0 \cdot \cdots \cdot \lambda_{i-1}}{\mu_1 \cdot \cdots \cdot \mu_i}\pi_0, \quad \text{where } \pi_0 \text{ is such that } \boldsymbol{\pi} \text{ sums to 1.} \tag{4}$$

We also have that the flow rate is:

$$\lambda^* = \sum_{i=0}^{K-1} \pi_i \lambda_i = \sum_{i=1}^{K} \pi_i \mu_i \tag{5}$$

We shall also be interested in systems for which $\lambda_0 \geq \cdots \geq \lambda_{K-1}$ and $\mu_1 \leq \cdots \leq \mu_K$. Examples include the M/M/c/K queue where service effort is increased when more customers are present, as well as systems where queue build up discourages arrivals.

Traffic processes In addition to the output process $\mathcal{D}$, we shall also be interested in the following counting processes: Let $\mathcal{A} = \{A(t), t \geq 0\}$ count arrivals, $\mathcal{E} = \{E(t), t \geq 0\}$ count entrances (admissions) and $\mathcal{L} = \{L(t), t \geq 0\}$ count overflows (jobs that arrive to a full system and are thus immediately lost). Immediate relations are:

$$A(t) = E(t) + L(t) \tag{6}$$

$$E(t) = Q(t) + D(t) \tag{7}$$

We shall also make use of the process $\mathcal{M} = \{M(t), t \geq 0\}$ defined as follows:

$$M(t) := E(t) + D(t) \tag{8}$$

$\mathcal{M}$ counts the number of transitions in the birth-death state space. The asymptotic variance rates of the processes $\mathcal{Q}, \mathcal{A}, \mathcal{E}, \mathcal{L}$ and $\mathcal{M}$ are defined similarly to $\overline{V}_{\mathcal{D}}$ (see (2)) and are labeled $\overline{V}_{\mathcal{Q}}, \overline{V}_{\mathcal{A}}, \overline{V}_{\mathcal{E}}, \overline{V}_{\mathcal{L}}$ and $\overline{V}_{\mathcal{M}}$ respectively. Note that when $\mathcal{A}$ is Poisson with rate λ, $\overline{V}_{\mathcal{A}} = \lambda$, and that $\overline{V}_{\mathcal{Q}} = 0$ because $0 \leq Q(t) \leq K$.

The following result shows that analysis of the entrances, outputs or transitions in terms of the asymptotic variance rate is equivalent:

Lemma 2.1 $\overline{V}_{\mathcal{E}} = \overline{V}_{\mathcal{D}} = \frac{1}{4}\overline{V}_{\mathcal{M}}$

Proof Using (7) we have, $\overline{V}_{\mathcal{E}} = \overline{V}_{\mathcal{Q}} + \overline{V}_{\mathcal{D}} + 2\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}}$ where,

$$\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}} := \lim_{t\to\infty} \frac{\mathrm{Cov}(Q(t), D(t))}{t},$$

is the asymptotic covariance rate of the pair $(\mathcal{Q},\mathcal{D})$. Using (8) and (7) we have $M(t) = Q(t) + 2D(t)$ and thus

$$\overline{V}_{\mathcal{M}} = \overline{V}_{\mathcal{Q}} + 4\overline{V}_{\mathcal{D}} + 4\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}}.$$

The result follows since $\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}}$ and $\overline{V}_{\mathcal{Q}}$ are 0. To show that $\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}} = 0$ we note:

$$\left| \frac{\mathrm{Cov}(Q(t), D(t))}{\sqrt{\mathrm{Var}(Q(t))(\overline{V}_{\mathcal{D}} t + o(t))}} \right| \leq 1$$

which implies that $\mathrm{Cov}(Q(t), D(t)) = O(\sqrt{t})$, and hence $\overline{\mathrm{Cov}}_{\mathcal{Q},\mathcal{D}} = 0$. □

MAPs We now briefly review Markov Arrival Processes (MAPs) and define the specific MAPs that are used throughout this paper. A brief description of MAPs is in [1], Chapter XI, Section 1a, more examples, results and applications are in [5] and [20]. A MAP, $\mathcal{N} = \{N(t), t \geq 0\}$, is a counting process specified by a generator matrix, $\mathbf{Q}$, of a finite irreducible CTMC on the states $\{0, \ldots, K\}$ with stationary distribution $\boldsymbol{\eta}$ (row vector), and two matrices, $\mathbf{C}$, $\mathbf{D}$ such that $\mathbf{Q} = \mathbf{C} + \mathbf{D}$. $\mathbf{C}$ has negative diagonal elements and non-negative off-diagonal elements. $\mathbf{D}$ is a non-negative matrix. In this paper we refer to $\mathbf{D}$ by the name: *event intensity matrix*.[1]

[1] Note that in other texts, the term "arrival" is generally used to refer to events because MAPs are often used to model arrival processes. Here we use "event" to avoid confusion.

$\mathcal{N}$ evolves as follows (loosely stated): When a CTMC (with generator $\mathbf{Q}$) makes a transition from state i to state j at time t, $N(t)$ is incremented w.p. d_{ij}/q_{ij}. $N(t)$ may also be incremented at times epochs during which the CTMC does not change state. This occurs according to a Poisson process with state dependent rate: d_{ii}. Thus the non-diagonal elements of $\mathbf{D}$ specify the proportion of transitions that are to be counted and the diagonal elements, allow to increase $\mathcal{N}$ by a Poisson process that is modulated by the state of the CTMC.

We assume that $\mathcal{N}$ has stationary increments, which occurs when the initial distribution of the underlying CTMC, with the generator $\mathbf{Q}$, is $\boldsymbol{\eta}$. The following results are summarized in [1]:

$$\mathbb{E}\,[N(t)] = \boldsymbol{\eta}\mathbf{D}\mathbf{1}\,t \tag{9}$$

$$\mathrm{Var}(N(t)) = \{\boldsymbol{\eta}\mathbf{D}\mathbf{1} - 2(\boldsymbol{\eta}\mathbf{D}\mathbf{1})^2 - 2\boldsymbol{\eta}\mathbf{D}\mathbf{Q}^{-}\mathbf{D}\mathbf{1}\}t + 2\boldsymbol{\eta}\mathbf{D}\mathbf{Q}^{-}(e^{\mathbf{Q}t} - \mathbf{I})\mathbf{Q}^{-}\mathbf{D}\mathbf{1} \tag{10}$$

where $\mathbf{Q}^{-} = (\mathbf{Q} - \mathbf{1}\boldsymbol{\eta})^{-1}$ and $\mathbf{I}$ is the identity matrix. We may express $\mathrm{Var}(N(t))$ without the matrix exponential as:

$$\mathrm{Var}(N(t)) = \overline{V}_{\mathcal{N}}\,t + \overline{B}_{\mathcal{N}} + O(t^{3r+2}e^{-bt})$$

for some integer r and $b > 0$ (cf. [1]). Here the asymptotic variance rate, $\overline{V}_{\mathcal{N}}$, and the y-intercept of the linear asymptote, $\overline{B}_{\mathcal{N}}$, are given by:

$$\overline{V}_{\mathcal{N}} = \boldsymbol{\eta}\mathbf{D}\mathbf{1} - 2(\boldsymbol{\eta}\mathbf{D}\mathbf{1})^2 - 2\boldsymbol{\eta}\mathbf{D}\mathbf{Q}^{-}\mathbf{D}\mathbf{1} \tag{11}$$

$$\overline{B}_{\mathcal{N}} = 2(\boldsymbol{\eta}\mathbf{D}\mathbf{1})^2 - 2\boldsymbol{\eta}\mathbf{D}\mathbf{Q}^{-}\mathbf{Q}^{-}\mathbf{D}\mathbf{1} \tag{12}$$

Clearly $\mathcal{M}$ and $\mathcal{D}$ are MAPs. With the exception of the numerical results of Sect. 5, all of the MAPs that we use have the birth-death generator matrix $\boldsymbol{\Lambda}$ as in (3). This implies that the event intensity matrix is all that is required to specify a MAP. The event intensity matrices for $\mathcal{D}$ and $\mathcal{M}$ are:

$$\mathbf{D}_{\mathcal{D}} = \begin{pmatrix} 0 & 0 & & & 0 \\ \mu_1 & \ddots & 0 & & \\ & \mu_2 & \ddots & \ddots & \\ & & \ddots & \ddots & 0 \\ 0 & & & \mu_K & 0 \end{pmatrix} \tag{13}$$

$$\mathbf{D}_{\mathcal{M}} = \begin{pmatrix} 0 & \lambda_0 & & & 0 \\ \mu_1 & \ddots & \lambda_1 & & \\ & \mu_2 & \ddots & \ddots & \\ & & \ddots & \ddots & \lambda_{K-1} \\ 0 & & & \mu_K & 0 \end{pmatrix} \tag{14}$$

It is easily verified that $\mathbb{E}\,[D(t)] = \lambda^* t$ and $\mathbb{E}\,[M(t)] = 2\lambda^* t$.

Fully counting MAPs and Markov modulated Poisson processes We define *Fully Counting MAPs* as MAPs for which the event intensity matrix consists of all the off diagonal elements of the generator Q, i.e. $\mathbf{D} = \mathbf{Q} - \mathrm{diag}(\mathbf{Q})$, where $\mathrm{diag}(\mathbf{Q})$ is a diagonal matrix with the same diagonal as $\mathbf{Q}$. In a fully counting MAP, all the events are state transitions of the underlying CTMC and every state transition of the underlying CTMC is an event, so that $N(t)$ is the number of all the transitions of the underlying stationary CTMC with generator $\mathbf{Q}$, over the period $[0, t]$. Note that $\mathcal{M}$ is a fully counting MAP but $\mathcal{D}$ is not.

When the event intensity matrix $\mathbf{D}$ of a MAP is a diagonal matrix then the MAP is a Markov Modulated Poisson Process (MMPP). All the events of a MMPP are generated by a doubly stochastic Poisson process whose rate is a function of the state of the underlying CTMC. A comprehensive reference about MMPPs is [13].

Fully counting MAPs and MMPPs are in a sense the extreme cases of MAPs. In a MMPP, the events do not coincide with state transitions (with probability 1). In contrast, in a fully counting MAP the events are precisely all the transitions of the CTMC. An early reference that analyzes both fully counting MAPs and MMPPs is [28].

Birth-death MMPPs Let $\tilde{\mathcal{N}}$ be a MMPP, with generator $\mathbf{Q}$ having stationary distribution $\boldsymbol{\eta}$. Denote the i'th diagonal element of $\mathbf{D}$ by $r(i)$ or r_i. This is the rate of events given that the CTMC is in state i. In Example 9.6.2 of [35], Whitt shows:

$$\overline{V}_{\tilde{\mathcal{N}}} = \sum_{i=0}^{K} r_i \eta_i + \overline{V}_{\mathcal{R}} \tag{15}$$

Here the asymptotic variance rate of the MMPP $\tilde{\mathcal{N}}$, $\overline{V}_{\tilde{\mathcal{N}}}$, is decomposed into two parts, where the first part is the average of the Poisson rate and the second part, $\overline{V}_{\mathcal{R}}$, is the asymptotic variance rate of the integrated rate process, $R(t) = \int_0^t r(Q(s))ds$. The Internet supplement of [35] shows how $\overline{V}_{\mathcal{R}}$ may be found from Poisson's equation for the CTMC (Theorem 2.3.4). In general this requires solving a system of linear equations (numerically), but when $\mathbf{Q}$ is birth-death, the following result holds (cf. [34], formula (6)):

$$\overline{V}_{\mathcal{R}} = 2 \sum_{i=0}^{K-1} \frac{1}{\eta_i \lambda_i} \left[\sum_{j=0}^{i} \left(r_j - \sum_{l=0}^{K} r_l \eta_l \right) \eta_j \right]^2 \tag{16}$$

We summarize (15) and (16) of Whitt as a proposition. It is one of the ingredients that yield the main result of this paper:

Proposition 2.2 *Let $\tilde{\mathcal{N}}$ be a MMPP with a birth-death generator matrix $\mathbf{Q}$ having birth rates λ_i, $i = 0, \ldots, K-1$ and stationary distribution η_i, $i = 0, \ldots, K$. Denote the diagonal elements of the event intensity matrix of $\tilde{\mathcal{N}}$ by r_i, $i = 0, \ldots, K$. Then the asymptotic variance of $\tilde{\mathcal{N}}$ is*:

$$\overline{V}_{\tilde{\mathcal{N}}} = \sum_{l=0}^{K} r_l \eta_l + 2 \sum_{i=0}^{K-1} \frac{1}{\eta_i \lambda_i} \left[\sum_{j=0}^{i} \left(r_j - \sum_{l=0}^{K} r_l \eta_l \right) \eta_j \right]^2.$$

3 Asymptotic variance rate of birth-death queues

We now consider a birth and death queue with generator Λ, stationary distribution $\boldsymbol{\pi}$ and flow rate λ^*, as in (3–5). We introduce the following notations, for $i = 0, \ldots, K-1$:

$$
\begin{aligned}
d_i &:= \lambda_i \pi_i, \\
D_i &:= \sum_{j=0}^{i} d_j \quad \text{(note that } D_{K-1} = \lambda^*\text{)}, \\
P_i &:= \sum_{j=0}^{i} \pi_j, \\
M_i &:= D_{i-1} - \lambda^* P_i \quad \text{(where we let } D_{-1} := 0\text{)}, \\
v_i &:= 2\left(M_i + \frac{M_i^2}{d_i}\right).
\end{aligned}
$$

Note that by the detailed balance equations, $d_i = \mu_{i+1}\pi_{i+1}$. Thus $D_{i-1} = \sum_{j=1}^{i} \mu_j \pi_j$, and hence M_i measures the difference between the actual rate of outputs observed on the states $\{0, 1, \ldots, i\}$ and the rate of outputs that would have been observed if the output rate on these states was uniformly equal to the flow rate, λ^*, independent of the state. Our main result is:

Theorem 3.1 *Let $\mathcal{Q}$ be a stationary CTMC with a birth-death structure as defined in* (3–5).

(i)

$$\overline{V}_{\mathcal{D}} = \lambda^* + \sum_{i=0}^{K-1} v_i$$

(ii) *If the birth and death rates of $\mathcal{Q}$ satisfy $\lambda_0 \geq \cdots \geq \lambda_{K-1}$ and $\mu_1 \leq \cdots \leq \mu_K$, then $v_i < 0$ for $i = 0, \ldots, K-1$ and as a result $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*} < 1$.*

Example 3.1 We may verify Theorem 3.1 for the M/M/1/1 queue (this example is also analyzed in [6]). This is a 2 state CTMC and it is the only M/M/1/K queue that has a renewal output process (cf. [11]). The distribution of the inter-output times is the convolution of an exponential rate λ and an exponential rate μ distribution. Thus the expectation is $m = \frac{1}{\lambda} + \frac{1}{\mu}$, the variance is $\sigma^2 = \frac{1}{\lambda^2} + \frac{1}{\mu^2}$ and since $\mathcal{D}$ is a renewal process,

$$\overline{V}_{\mathcal{D}} = \frac{\sigma^2}{m^3} = \frac{\lambda\mu(\lambda^2 + \mu^2)}{(\lambda + \mu)^3}$$

Now, $P_0 = \pi_0 = \frac{\mu}{\lambda+\mu}$, $\lambda^* = d_0 = \frac{\lambda\mu}{\lambda+\mu}$, $M_0 = -\frac{\lambda\mu^2}{(\lambda+\mu)^2}$ and $v_0 = -2\frac{\lambda^2\mu^2}{(\lambda+\mu)^3}$ (notice it is negative). And we obtain $\lambda^* + v_0 = \frac{\sigma^2}{m^3}$.

To prove Theorem 3.1 we use the following result, which is also of independent interest.

Proposition 3.2 *Let* $\mathbf{Q}$ *be a generator of a finite state irreducible CTMC. For any* $0 \le \alpha \le 1$ *let* $\mathcal{N}_\alpha = \{N_\alpha(t), t \ge 0\}$ *be a stationary MAP with generator* $\mathbf{Q}$ *and event intensity matrix* $\mathbf{D}_\alpha = \alpha\mathbf{Q} - \mathrm{diag}(\mathbf{Q})$. *Then* $\mathbb{E}[N_\alpha(t)]$ *and* $Var(N_\alpha(t))$ *are independent of* α.

Note that when $\alpha = 1$ we have a fully counting MAP and when $\alpha = 0$ we have a MMPP.

Proof From (9) and (10) we see that $\mathbb{E}[N_\alpha(t)]$ and $\mathrm{Var}(N_\alpha(t))$ only depend on $\mathbf{D}_\alpha\mathbf{1}$ and $\boldsymbol{\eta}\mathbf{D}_\alpha$.

First observe that $\mathbf{D}_\alpha\mathbf{1}$ is independent of α: Denote the elements of $\mathbf{Q}$ by q_{ij}. $\mathbf{Q}$ is a generator matrix so $q_{ii} = -\sum_{j\ne i} q_{ij}$, thus the i'th element of $\mathbf{D}_\alpha\mathbf{1}$ is:

$$\alpha \sum_{j\ne i} q_{ij} - (1-\alpha)q_{ii} = \sum_{j\ne i} q_{ij}$$

Next observe that $\boldsymbol{\eta}\mathbf{D}_\alpha$ is independent of α: Since $\boldsymbol{\eta}$ is the stationary distribution we have $\boldsymbol{\eta}\mathbf{Q} = \mathbf{0}$. Thus $\boldsymbol{\eta}\mathbf{D}_\alpha = \alpha\boldsymbol{\eta}\mathbf{Q} - \boldsymbol{\eta}\mathrm{diag}(\mathbf{Q}) = -\boldsymbol{\eta}\mathrm{diag}(\mathbf{Q})$. □

We are now ready to prove Theorem 3.1:

Proof of Theorem 3.1 (i) Let $\overline{V}_{\tilde{\mathcal{M}}}$ be the asymptotic variance rate of the MAP (also MMPP) $\tilde{\mathcal{M}} = \{\tilde{M}(t), t \ge 0\}$ having the following event intensity matrix:

$$\mathbf{D}_{\tilde{\mathcal{M}}} = \begin{pmatrix} \lambda_0 & & & & & 0 \\ & \mu_1+\lambda_1 & & & & \\ & & \mu_2+\lambda_2 & & & \\ & & & \ddots & & \\ & & & & \mu_{K-1}+\lambda_{K-1} & \\ 0 & & & & & \mu_K \end{pmatrix} \tag{17}$$

Denote the diagonal elements of $\mathbf{D}_{\tilde{\mathcal{M}}}$ by r_i, $i = 0, \ldots, K$. We now have:

$$4\overline{V}_{\mathcal{D}} = \overline{V}_{\mathcal{M}} = \overline{V}_{\tilde{\mathcal{M}}} = \sum_{l=0}^{K} r_l\pi_l + 2\sum_{i=0}^{K-1} \frac{1}{\pi_i\lambda_i}\left[\sum_{j=0}^{i}\left(r_j - \sum_{l=0}^{K} r_l\pi_l\right)\pi_j\right]^2 \tag{18}$$

The first equality is from Lemma 2.1. The second equality is from Proposition 3.2 by taking $\alpha = 1$ for the fully counting MAP, $\mathcal{M}$ and $\alpha = 0$ for the MMPP, $\tilde{\mathcal{M}}$ (see (14) and (17)). The third equality is from Proposition 2.2 since $\tilde{\mathcal{M}}$ is a Birth and Death

MMPP. Now note that $\sum_{l=0}^{K} \pi_l r_l = 2\lambda^*$ and

$$\sum_{j=0}^{i}\left(r_j - \sum_{l=0}^{K} r_l \pi_l\right)\pi_j = \sum_{j=0}^{i} \lambda_j \pi_j + \sum_{j=1}^{i} \mu_j \pi_j - 2\lambda^* \sum_{j=0}^{i} \pi_j = d_i + 2(D_{i-1} - \lambda^* P_i) \tag{19}$$

The first equality follows from direct substitution of r_j and the second follows from the detailed balance equations $\mu_i \pi_i = \lambda_{i-1}\pi_{i-1}$ and simplification. Now using the definition of M_i and substituting (19) in (18) we get:

$$4\overline{V}_{\mathcal{D}} = 2\lambda^* + 2\sum_{i=0}^{K-1} \frac{d_i^2 + 4d_i M_i + 4M_i^2}{d_i}$$

Noticing that $\sum_{i=0}^{K-1} d_i = \lambda^*$, result (i) follows.

(ii) We use the following two simple inequalities:

(a) For $a, b, c, d > 0$

$$\frac{a}{b} < \frac{c}{d} \quad \Leftrightarrow \quad \frac{a}{b} < \frac{a+c}{b+d} \quad \Leftrightarrow \quad \frac{a+c}{b+d} < \frac{c}{d}$$

(b) For $a, b, c, d, \Delta > 0$

$$\frac{a}{b} \le \frac{c}{d} \quad \text{and} \quad a < b \quad \Rightarrow \quad \frac{a}{\Delta + b} < \frac{c}{\Delta + d}$$

From $\lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_{K-1}$ we get:

$$\frac{d_0}{\pi_0} \ge \frac{d_1}{\pi_1} \ge \cdots \ge \frac{d_{K-1}}{\pi_{K-1}}$$

and therefore using (a):

$$\frac{D_0}{P_0} \ge \frac{D_1}{P_1} \ge \cdots \ge \frac{D_{K-1}}{P_{K-1}} > D_{K-1} = \lambda^*$$

From $\mu_1 \le \mu_2 \le \cdots \le \mu_K$ we get:

$$\frac{d_0}{\pi_1} \le \frac{d_1}{\pi_2} \le \cdots \le \frac{d_{K-1}}{\pi_K}$$

and therefore using (b):

$$\frac{D_0}{P_1 - \pi_0} \le \frac{D_1}{P_2 - \pi_0} \le \cdots \le \frac{D_{K-1}}{P_K - \pi_0}$$

and furthermore, since $D_0 < D_1 < \cdots < D_{K-1}$ we also have:

$$0 < \frac{D_0}{P_1} < \frac{D_1}{P_2} < \cdots < \frac{D_{K-1}}{P_K} = \lambda^*.$$

Hence, for all $i = 0, \ldots, K-1$,

$$d_i + D_{i-1} - \lambda^* P_i > 0 > D_{i-1} - \lambda^* P_i,$$

which implies:

$$M_i < 0 \text{ and } d_i > |M_i|$$

from which it follows that:

$$v_i = 2(M_i + \frac{M_i^2}{d_i}) < 0$$

(to clarify: $d_i + M_i > 0 > M_i \Rightarrow d_i > -M_i > 0 \Rightarrow d_i|M_i| > M_i^2 \Rightarrow -M_i = |M_i| > \frac{M_i^2}{d_i} \Rightarrow 0 > M_i + \frac{M_i^2}{d_i}$). □

Note that the condition on the birth and death rates in Part (ii) implies that the sequence $\frac{\lambda_i}{\mu_{i+1}}$, $i = 0, \ldots, K-1$ is non-increasing and as a result $\boldsymbol{\pi}$ is unimodal (cf. [18]). This observation makes it tempting to attempt to generalize the theorem in this direction. The following example shows that this is not possible:

Example 3.2 Let $K = 2$, $\lambda_0 = \frac{1}{3}$, $\mu_1 = \frac{1}{3}$, $\lambda_1 = 1$ and $\mu_2 = \frac{3}{2}$. The stationary distribution is $(\pi_0\ \pi_1\ \pi_2) = (\frac{3}{8}\ \frac{3}{8}\ \frac{1}{4})$. It is unimodal as expected because $\frac{\lambda_i}{\mu_{i+1}}$ is non-increasing but $v_0 = \frac{3}{16}$, $v_1 = -\frac{1}{6}$ and $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*} = 1 + \frac{1}{24}$.

4 The M/M/1/K queue

We now apply Theorem 3.1 to the case where the birth and death rates are constant, $\lambda, \mu > 0$. Denote $\rho = \frac{\lambda}{\mu}$. The stationary distribution and the flow rate are:

$$\begin{aligned} \pi_i &= \begin{cases} \frac{1}{K+1} & \rho = 1 \\ \rho^i \frac{1-\rho}{1-\rho^{K+1}} & \rho \neq 1 \end{cases} \qquad i = 0, \ldots, K \\ \lambda^* &= \begin{cases} \lambda \frac{K}{K+1} & \rho = 1 \\ \lambda \frac{1-\rho^K}{1-\rho^{K+1}} & \rho \neq 1 \end{cases} \end{aligned} \tag{20}$$

Corollary 4.1 *For the $M/M/1/K$ queue*:

$$\begin{aligned} \overline{V}_{\mathcal{D}} &= \begin{cases} \lambda \frac{2K^2+K}{3K^2+6K+3} & \rho = 1 \\ \lambda \frac{(1+\rho^{K+1})(1-(1+2K)\rho^K(1-\rho)-\rho^{2K+1})}{(1-\rho^{K+1})^3} & \rho \neq 1 \end{cases} \\ \frac{\overline{V}_{\mathcal{D}}}{\lambda^*} &= \begin{cases} \frac{2K+1}{3K+3} & \rho = 1 \\ \frac{(1+\rho^{K+1})(1-(1+2K)\rho^K(1-\rho)-\rho^{2K+1})}{(1-\rho^K)(1-\rho^{K+1})^2} & \rho \neq 1 \end{cases} \end{aligned} \tag{21}$$

Proof Using straight forward (but lengthy) calculations we obtain:

$$M_i = \begin{cases} -\lambda \frac{K-i}{(K+1)^2} & \rho = 1 \\ -\lambda \rho^i \frac{(1-\rho)(1-\rho^{K-i})}{(1-\rho^{K+1})^2} & \rho \neq 1 \end{cases} \qquad i = 0, \ldots, K-1$$

$$v_i = \begin{cases} -\lambda 2 \frac{(i+1)(K-i)}{(K+1)^3} & \rho = 1 \\ -\lambda 2 \rho^K \frac{(1-\rho^{i+1})(1-\rho)(1-\rho^{K-i})}{(1-\rho^{K+1})^3} & \rho \neq 1 \end{cases} \qquad i = 0, \ldots, K-1$$

The result follows from Theorem 3.1 after summation of finite geometric series and simplification. □

The following properties of $\overline{V}_{\mathcal{D}}$ and $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*}$ should be noted:

- For fixed K, $\overline{V}_{\mathcal{D}}$ and $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*}$ are continuous in λ and μ for all $\lambda,\ \mu > 0$.
- For fixed $\lambda,\ \mu$ we have:

$$\lim_{K\to\infty} \overline{V}_{\mathcal{D}} = \begin{cases} \lambda & \lambda < \mu \\ \frac{2}{3}\lambda & \lambda = \mu \\ \mu & \lambda > \mu \end{cases}$$

- For fixed K and fixed $C > 0$, $\overline{V}_{\mathcal{D}}$ and $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*}$ are symmetric about the point $\lambda = \mu$ on the interval $\{(\lambda, \mu) \mid \lambda + \mu = C,\ \lambda,\ \mu > 0\}$ (see also Fig. 2).

The following corollary formalizes the BRAVO effect for M/M/1/K:

Corollary 4.2 *Consider the $M/M/1/K$ queue with $\lambda + \mu = C$ for some $C > 0$. Then when $\lambda = \mu$: $\overline{V}_{\mathcal{D}}$ is locally minimized and $\frac{\overline{V}_{\mathcal{D}}}{\lambda^*}$ is globally minimized.*

Proof Take derivatives and limits of the expressions of Corollary 4.1. □

4.1 Asymptotic correlation between outputs and overflows

It is well known and easy to observe that the overflow process, $\mathcal{L}$, is a renewal process. The overflow rate is of course $\lambda - \lambda^*$. Berger and Whitt, [3] in their (6) derive the SCV for the inter-overflow times. Multiplying these we obtain the asymptotic variance rate of the overflows:[2]

$$\overline{V}_{\mathcal{L}} = \begin{cases} \lambda \frac{2K^2+4K+3}{3K^2+6K+3} & \rho = 1 \\ \lambda \frac{(\rho^K - \rho^{3K+2})(1+\rho) - 4(K+1)(1-\rho)\rho^{2K+1}}{(1-\rho^{K+1})^3} & \rho \neq 1 \end{cases} \tag{22}$$

[2] An alternative derivation of (22) is by conditioning $L(t)$ on the occupation time of state K during $[0, t]$, and using the conditional variance formula. This calculation requires evaluation of the asymptotic variance rate of the occupation time using formula (16).

In general, the covariance, asymptotic covariance rate, correlation and limiting correlation of pairs of traffic processes may be numerically calculated by modeling the queueing system as a Marked Markovian Arrival Process (MMAP) (cf. [15]) and using formulas similar to (10) that appear in that reference. For the simple case of the M/M/1/1 queue, an explicit expression was obtained for the limiting correlation coefficient between the outputs and the overflows in [6]. We now extend this result:

Corollary 4.3 *For the M/M/1/K queue*:

$$\lim_{t\to\infty}\operatorname{Corr}(E(t),L(t))=\lim_{t\to\infty}\operatorname{Corr}(D(t),L(t))=\begin{cases}\overline{R}_{\rho,K} & \rho<1\\ -\frac{1-\frac{1}{K}}{4\sqrt{1+\frac{5}{2K}+\frac{5}{2K^2}+\frac{3}{4K^3}}} & \rho=1\\ -\overline{R}_{\rho,K} & \rho>1\end{cases}$$

where:

$$\overline{R}_{\rho,K}=\frac{\rho^{\frac{K}{2}}(K(1-\rho)(1+3\rho^{1+K})-\rho(1-\rho^K)(3+\rho^{K+1}))}{\sqrt{(1+\rho^{K+1})(1-(2K+1)(1-\rho)\rho^K-\rho^{2K+1}))((1+\rho)(1-\rho^{2K+2})-4(K+1)(1-\rho)\rho^{K+1})}}$$

Proof In a similar manner to the proof of Lemma 2.1, define the asymptotic covariance rates $\overline{\operatorname{Cov}}_{\mathcal{D},\mathcal{L}}$ and $\overline{\operatorname{Cov}}_{\mathcal{E},\mathcal{L}}$. We are assured that the covariance functions of these traffic process are $O(t)$ since $\operatorname{Var}(D(t))$, $\operatorname{Var}(E(t))$ and $\operatorname{Var}(L(t))$ are $O(t)$.

Take the variance of (6), divide by t, take the limit $t\to\infty$ and rearrange to arrive at:

$$\overline{\operatorname{Cov}}_{\mathcal{E},\mathcal{L}}=\frac{\lambda-\overline{V}_{\mathcal{E}}-\overline{V}_{\mathcal{L}}}{2}$$

In a similar manner (and using arguments similar to the proof of Lemma 2.1), obtain:

$$\overline{\operatorname{Cov}}_{\mathcal{D},\mathcal{L}}=\frac{\lambda-\overline{V}_{\mathcal{D}}-\overline{V}_{\mathcal{L}}}{2}$$

Thus from Lemma 2.1 and from substitution of (21), (22)

$$\overline{\operatorname{Cov}}_{\mathcal{E},\mathcal{L}}=\overline{\operatorname{Cov}}_{\mathcal{D},\mathcal{L}}=\begin{cases}-\lambda\frac{K^2-K}{6K^2+12K+6} & \rho=1\\ -\lambda\frac{(1-\rho^K)(3+\rho^{K+1})\rho^{K+1}-K(1-\rho)(1+3\rho^{K+1})\rho^K}{(1-\rho^{K+1})^3} & \rho\neq1\end{cases}\tag{23}$$

The correlation coefficient is obtained directly from (21), (22), (23) and simplification. □

Figure 3 displays the limiting correlation coefficient for various buffer sizes. Note also the following properties:

- The limiting correlation is continuous in ρ for all $\rho>0$.
- For fixed K, as $\rho\to\infty$, the limiting correlation increases to 0.

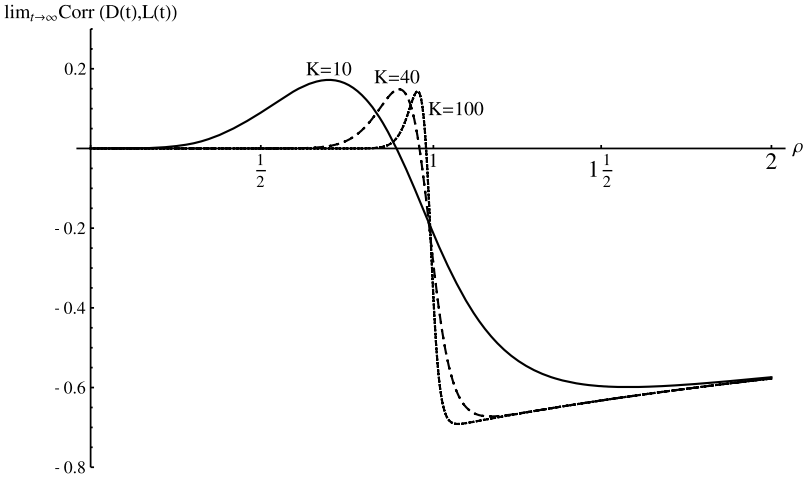

Fig. 3 M/M/1/K: The limiting correlation between entrances/outputs and overflows as a function of ρ

- For fixed ρ we have:

$$\lim_{K\to\infty}\lim_{t\to\infty}\mathrm{Corr}(D(t),L(t)) = \begin{cases} 0 & \rho<1 \\ -\frac{1}{4} & \rho=1 \\ -\frac{1}{\sqrt{1+\rho}} & \rho>1 \end{cases}$$

- For finite K, let $\hat{\rho} := \arg\max_{0<\rho<1} \overline{R}_{\rho,K}$. Then $\hat{\rho}$ converges to 1 as $K\to\infty$, and it is numerically observed that the maximum value converges to $\lim_{K\to\infty}\overline{R}_{\hat{\rho},K} \approx 0.139772$.
- Similarly, let $\breve{\rho} := \arg\min_{\rho>1} -\overline{R}_{\rho,K}$. Then $\breve{\rho}$ converges to 1 as $K\to\infty$, and the minimum value converges to $\lim_{K\to\infty} -\overline{R}_{\breve{\rho},K} = -\frac{1}{\sqrt{2}}$.
- Summarizing, informally, we see that the limiting correlation attains 3 different values at the vicinity of $\rho=1$ for large K:

$$\lim_{t\to\infty}\mathrm{Corr}(D(t),L(t)) \approx \begin{cases} 0.139772 & \rho=1^- \\ -\frac{1}{4} & \rho=1 \\ -\frac{1}{\sqrt{2}} & \rho=1^+ \end{cases}$$

4.2 The y-intercept of the linear asymptote of Var$(D(t))$

We now analyze the y-intercept, $\overline{B}_{\mathcal{D}}$ according to formula (12) (take $\mathbf{Q}=\mathbf{\Lambda}$, $\boldsymbol{\eta}=\boldsymbol{\pi}$ and $\mathbf{D}=\mathbf{D}_{\mathcal{D}}$). Figure 4 presents $\overline{B}_{\mathcal{D}}$ as a function of λ for $K=10$ and $K=20$. Interestingly, $\overline{B}_{\mathcal{D}}$ appears to be maximized when $\rho=1$ and the value increases with K. Note that when $\rho\neq 1$ our calculations indicate that $\overline{B}_{\mathcal{D}}$ decreases to 0 as $K\to\infty$.

The values of $\overline{B}_{\mathcal{D}}$ in Fig. 4 were evaluated numerically (each point on the curve requires inversion of $(\mathbf{\Lambda}-\mathbf{1}\boldsymbol{\pi})$ to obtain $\mathbf{\Lambda}^-$). In the balanced case the stationary

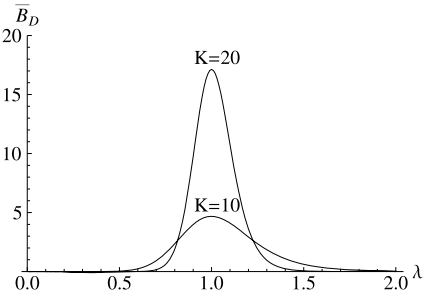

Fig. 4 M/M/1/K: $\overline{B}_{\mathcal{D}}$ as a function λ when $\mu = 1$

distribution of the queue lengths (20) is discrete uniform, and we can obtain explicit expressions for the elements of $\mathbf{\Lambda}^-$ and in turn have an explicit expression for $\overline{B}_{\mathcal{D}}$. The following proposition derives this expression and shows that $\overline{B}_{\mathcal{D}}$ for $\rho = 1$ increases quadratically with K.

Proposition 4.4 *For the $M/M/1/K$ queue with $\rho = 1$ and $K \geq 2$:*

$$\overline{B}_{\mathcal{D}} = \frac{7K^4 + 28K^3 + 37K^2 + 18K}{180K^2 + 360K + 180}$$

Proof For $\gamma > 0$ and $K \geq 2$, define the $K \times K$ matrix:

$$\mathbf{A}_K^{\gamma} = \begin{pmatrix} 1+\gamma & 1-\gamma & & & & & & 1 \\ 1-\gamma & 1+2\gamma & \ddots & & & & & \\ & \ddots & 1+2\gamma & \ddots & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & 1+2\gamma & \ddots & & \\ & & & & \ddots & 1+2\gamma & 1-\gamma & \\ & & & & & 1-\gamma & 1+\gamma & \\ 1 & & & & & & & \end{pmatrix}$$

Its inverse, $(\mathbf{A}_K^{\gamma})^{-1}$, is also a symmetric matrix with elements given by:

$$\tilde{a}_{i,j} = \frac{i^2 - i + (K+1-j)^2 - (K+1-j)}{2K\gamma} - \frac{K^3 - K - 6\gamma}{6K^2\gamma}, \quad i \leq j. \tag{24}$$

This can be verified by showing $\mathbf{A}_K^{\gamma}(\mathbf{A}_K^{\gamma})^{-1} = \mathbf{I}$. This is a lengthy but straight forward calculation, we omit the details.

Now observe that $\mathbf{\Lambda}^- = -(K+1)(\mathbf{A}_{K+1}^{\lambda(K+1)})^{-1}$ and find the elements of $\mathbf{\Lambda}^-\mathbf{\Lambda}^-$ by multiplication. These are rather complicated expressions and we again omit the details. We now have:

$$\begin{aligned} \overline{B}_{\mathcal{D}} &= 2\left(\lambda\frac{K}{K+1}\right)^2 - 2\frac{1}{K+1}\mathbf{1}'\mathbf{D}\mathbf{\Lambda}^-\mathbf{\Lambda}^-\mathbf{D}\mathbf{1} \\ &= 2\left(\lambda\frac{K}{K+1}\right)^2 - 2\frac{\lambda^2}{K+1}(1,\ldots,1,0)\mathbf{\Lambda}^-\mathbf{\Lambda}^-(0,1,\ldots,1)' \end{aligned} \tag{25}$$

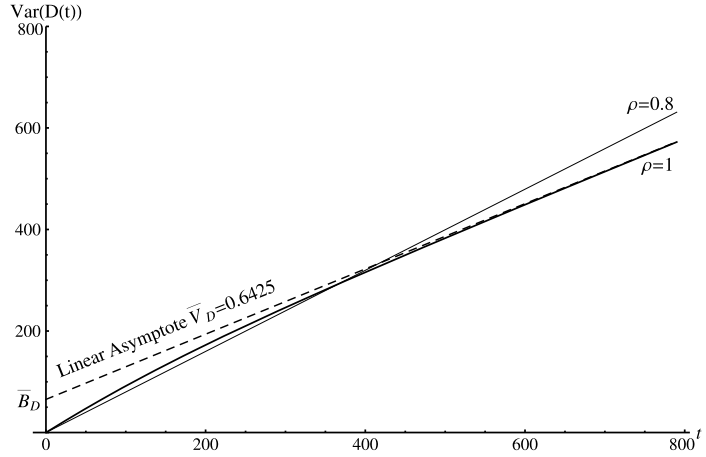

Fig. 5 M/M/1/40: Var($D(t)$) for $\mu = 1$ and two different arrival rates. Heavy curve is for $\lambda = 1$ (balanced). *Light curve* is for $\lambda = 0.8$ (unbalanced). *Dashed line* is linear asymptote of balanced case

The bilinear form in the second term is a summation of all entries of the matrix $\mathbf{\Lambda}^- \mathbf{\Lambda}^-$ except for the first column and last row. The resulting expression is:

$$(1, \ldots, 1, 0)\mathbf{\Lambda}^- \mathbf{\Lambda}^- (0, 1, \ldots, 1)' = -\frac{7K^4 + 28K^3 - (360\lambda^2 - 37)K^2 + 18K}{360\lambda^2(K+1)}$$

Plugging this in (25) and simplifying we obtain the result. □

4.3 Var($D(t)$) in the short range

We now present numerical examples and results of the variance function for finite t. While our main finding of this paper is that balancing reduces Var($D(t)$) in the long range (BRAVO), there is no guarantee that it has the same effect in the short range. In fact, Fig. 4 hints that balanced systems may have a higher variance function then unbalanced systems in the short range since the y-intercept of their linear asymptote is higher.

This is illustrated in Fig. 5. Here we compare the variance function, Var($D(t)$), of a balanced system to that of a system with $\rho = 0.8$. We plot the variance function (heavy curve) and its linear asymptote (dashed line) for the balanced system, and the variance function for the unbalanced system (light curve). Both are calculated for $K = 40$, using formula (10). It is observed that for the balanced system, the slope of the variance function is steeper than the asymptotic variance rate for small t and nears the asymptotic variance rate of approximately $\frac{2}{3}$ as time progresses. On the contrary the slope of the variance function of an unbalanced system almost equals the asymptotic variance (approximately 0.8, with negligible intercept) from the outset. As a result, the unbalanced system has a slightly lower variance function for values of t smaller than approximately 350.

To further understand the short-term behavior we performed extensive calculations in which we compared the variance of the output process of balanced M/M/1/K

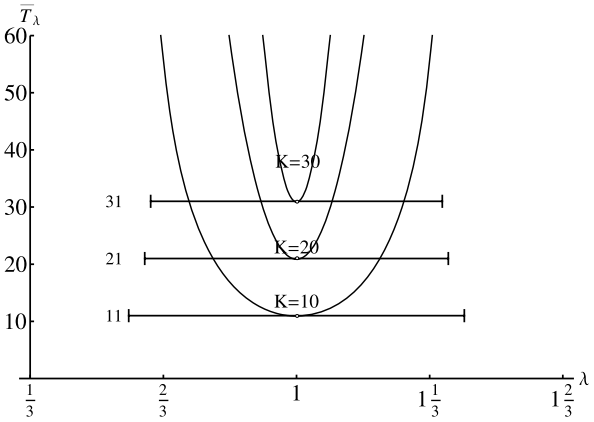

Fig. 6 $\overline{T}_\lambda$ for $K = 10, 20, 30$. *The horizontal intervals* are exactly at heights 11, 21, 31 showing that $\overline{T}_{1^-} = \overline{T}_{1^+} \approx K + 1$. *The horizontal intervals* also specify the range of values for which $\overline{V}_{\mathcal{D}_1} \leq \overline{V}_{\mathcal{D}_\lambda}$

queues, $\mathcal{D}_1 = \{D_1(t), t \geq 0\}$, to that of unbalanced queues, with arrival rates λ and service rates $\mu = 2 - \lambda$, given by $\mathcal{D}_\lambda = \{D_\lambda(t), t \geq 0\}$. We define:

$$\overline{T}_\lambda := \inf\{t > 0 \mid \mathrm{Var}(D_1(t)) \leq \mathrm{Var}(D_\lambda(t))\}$$

Stated informally, $\overline{T}_\lambda$ is a measure of the time it takes the BRAVO effect to "kick-in" when comparing a balanced system to an unbalanced one. We evaluated $\overline{T}_\lambda$ only for λ for which $\overline{V}_{\mathcal{D}_1} < \overline{V}_{\mathcal{D}_\lambda}$. It is infinite otherwise. The range of these λ's varies with K and as $K \to \infty$ the range converges to $(\frac{2}{3}, \frac{4}{3})$. Figure 6 shows $\overline{T}_\lambda$ as a function of λ for $K = 10, 20, 30$. We observe the following:

- $0 < \overline{T}_\lambda < \infty$ for all $\lambda \neq 1$ that satisfy $\overline{V}_{\mathcal{D}_1} < \overline{V}_{\mathcal{D}_\lambda}$. The fact that $\overline{T}_\lambda \neq 0$ shows that indeed the unbalanced systems have a lower variance in the short range (during the time interval $[0, \overline{T}_\lambda)$).
- For fixed $\lambda \neq 1$, $\overline{T}_\lambda$ increases with K. In fact, for λ far enough from 1, a simple approximation for $\overline{T}_\lambda$ may be achieved by calculating the intersection of the linear asymptote of the balanced system (it is given by Corollary 4.1 and Proposition 4.4) and an approximation of the linear asymptote of an unbalanced system taking the y-intercept to be 0 and the asymptotic variance rate to be λ. According to this approximation, $\overline{T}_\lambda$ increases quadratically with K.
- For $\lambda = 1^-$ and $\lambda = 1^+$ we observe $\overline{T}_\lambda \approx K + 1$ and the approximation quickly becomes accurate when K increases. Note that $\overline{T}_1$ is trivially 0 and thus there is a singularity in the function $\overline{T}_\lambda$ at $\lambda = 1$. We do not have any intuitive explanation for the value of $K + 1$ at the moment.

5 More on BRAVO

For the M/M/1/K queue, our intuition for BRAVO is as follows: Since the asymptotic variance rate of the transitions $\mathcal{M}$ and the outputs $\mathcal{D}$ are the same up to a constant we can gain intuition by considering the transitions process. Now it can be seen that the rate of transitions incurred on states $\{1, \ldots, K-1\}$ is $\lambda + \mu$ while the rates of

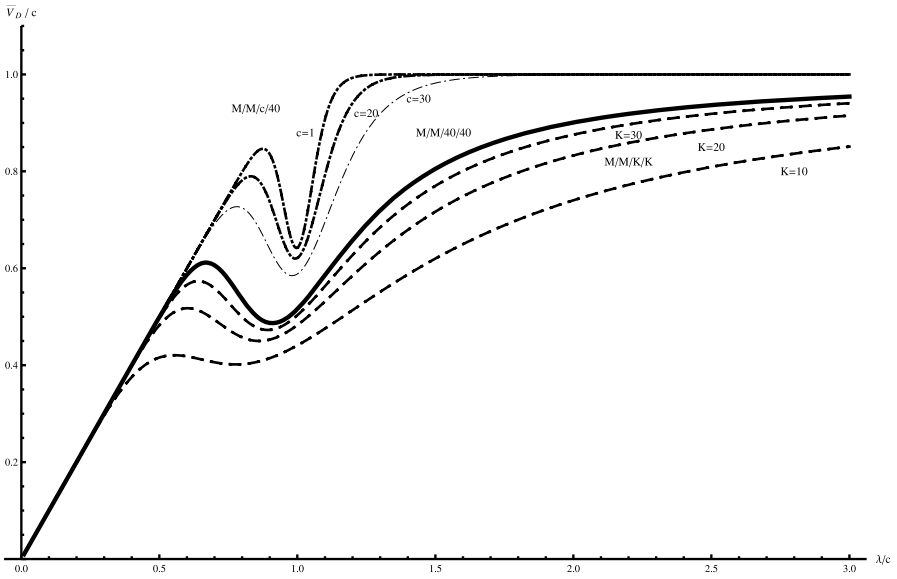

Fig. 7 M/M/c/K: $\frac{\overline{V}_{\mathcal{D}}}{c}$ as a function of $\frac{\lambda}{c}$ when $\mu = 1$

transitions on the edge states, 0 and K are λ and μ respectively. Observing the steady state distribution, (20), we see that when $\lambda = \mu$ the system spends very little time on the edge states and thus the "modulation" between rates $\lambda + \mu$ and λ or μ is minimal. On the contrary when $\lambda \neq \mu$ the system often switches between an edge state and a non-edge state and thus there is substantial "modulation" in the transition rates and as a result the variance of the transition process is greater.

This intuition does not immediately carry over to more complex systems but the BRAVO effect does. We now show some examples.

5.1 M/M/c/K

The M/M/c/K queue with $1 \leq c \leq K$ is an example a of a birth-death queue with monotone rates (the birth rates are constant and the death rates are increasing). While Theorem 3.1 is applicable to this system, the calculation of the normalization constant of the stationary distribution does not simplify and thus we are not able to obtain simple a formula for $\overline{V}_{\mathcal{D}}$ except for the case $c = 1$ (Sect. 4). Nevertheless, the computation of $\overline{V}_{\mathcal{D}}$ using the formula of (3.1) is simpler and more efficient than using the matrix formula (11).

Figure 7 shows that the BRAVO effect appears in the M/M/c/K queue: in this case "**b**alancing" implies setting $\lambda = c\mu$. The thick curve is for the Erlang loss system ($c = K$) with $K = 40$ to which we compare other systems. It is apparent that as the number of servers decreases, the asymptotic variance rate normalized by the number of servers increases. Alternatively, keeping the number of servers equal to the buffer size and decreasing the number of servers causes a decrease in the asymptotic

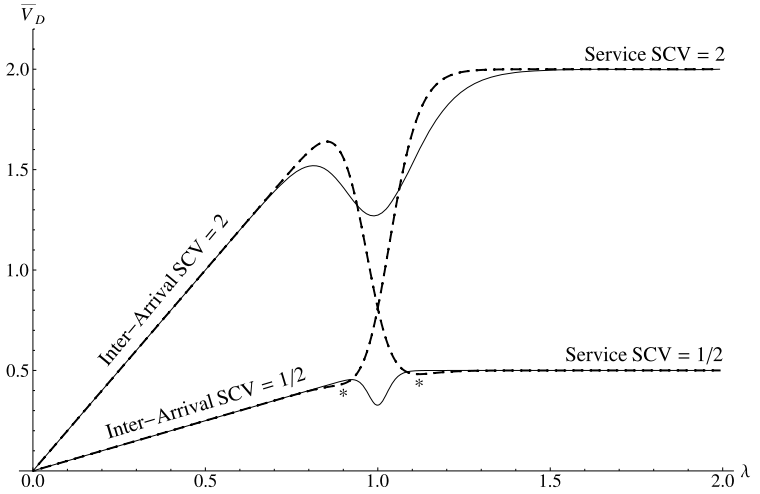

Fig. 8 PH/PH/1/40: $\overline{V}_{\mathcal{D}}$ as a function of λ when $\mu = 1$ for four combinations of inter-arrival and service times distributions

variance rate normalized by the number of servers. We do not yet have an intuitive explanation for BRAVO in the M/M/c/K queue.

5.2 Non exponential distributions

We now consider some examples of GI/G/1/K using phase-type distributions (cf. [5]). We let the inter-arrival and/or service time distributions be generated by sequences of i.i.d Erlang random variables, $\{E_1, E_2, \ldots\}$, and i.i.d hyper-exponential random variables, $\{H_1, H_2, \ldots\}$:

$$E_1 \sim \text{Erlang}(2, 2)$$

$$H_1 \sim \begin{cases} \exp(\frac{1}{2}) & \text{w.p. } 1/3 \\ \exp(2) & \text{w.p. } 2/3 \end{cases}$$

Note that: $\mathbb{E}[E_1] = \mathbb{E}[H_1] = 1$, the SCV of E_1 is $\frac{1}{2}$ and the SCV of H_1 is 2. We denote the queueing systems with the four possible combinations of inter-arrival and service distributions by: E/E/1/K, H/H/1/K, E/H/1/K and H/E/1/K. In all our examples we set $\mu = 1$ and scale the corresponding sequences of inter-arrival or service times by $\frac{1}{\lambda}$.

These are simple examples of PH/PH/1/K queues and are represented by a CTMC with $2 + 4K$ states (in this example, both E_1 and H_1 are PH distributions with 2 phases). Now using formula (11) for various values of λ we obtain Fig. 8. The solid curves are for the E/E/1/K and H/H/1/K cases. The dashed curves are for the E/H/1/K and H/E/1/K cases.

When $\lambda \gg \mu$ we expect the asymptotic variance rate to be determined by the service distribution. This is because the server is almost always busy and thus we almost have a renewal output process with asymptotic variance μc_S^2 (where c_S^2 is the SCV

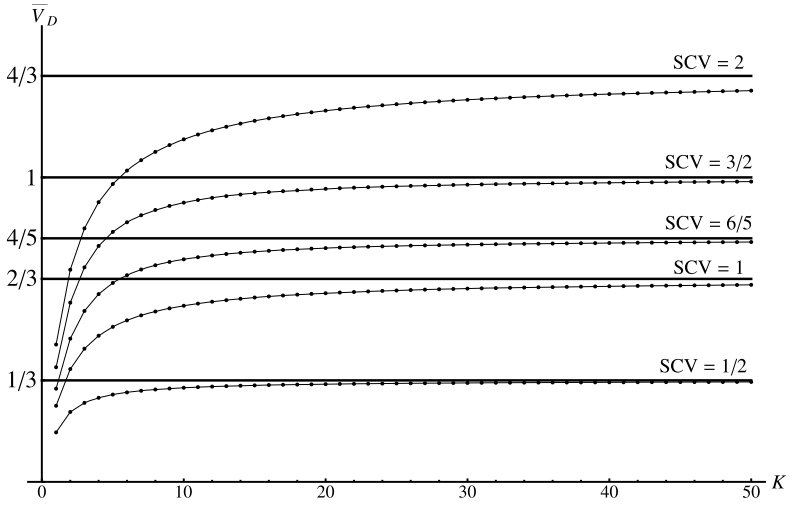

Fig. 9 PH/PH/1/K: $\overline{V}_{\mathcal{D}}$ as a function of K for $\lambda = \mu = 1$, for various values of the SCVs of inter-arrival and service times. The horizontal lines are at $\frac{2}{3}$SCV

of the service distribution). Similarly, when $\lambda \ll \mu$, we expect the asymptotic variance rate to be determined by the inter-arrival distribution. This is because the overflow rate is very small and thus $A(t) \approx Q(t) + D(t)$. Now, since $Q(t)$ is bounded, $\overline{V}_{\mathcal{D}} \approx \overline{V}_{\mathcal{A}} = \lambda c_A^2$ (where c_A^2 is the SCV of the inter-arrival distribution). Now consider the case where $\lambda \approx \mu$: In the E/E/1/K and H/H/1/K systems (same SCV for inter-arrival and service distributions) we clearly observe the BRAVO effect: these curves have a pronounced local minimum at the vicinity of $\rho = 1$.

In the E/H/1/K and H/E/1/K systems, we observe a "smoothed step" in $\overline{V}_{\mathcal{D}}$ at the vicinity of $\rho = 1$ between the values, 2 and $\frac{1}{2}$ (approximately for finite K). This is due to the fact that for $\lambda \ll \mu$ we have $\overline{V}_{\mathcal{D}} \approx \lambda c_A^2$ and for $\lambda \gg \mu$ we have $\overline{V}_{\mathcal{D}} \approx \mu c_S^2$ and $c_A^2 \neq c_S^2$ and is not directly due to the BRAVO effect. Nevertheless, we believe that traces of the BRAVO effect appear in the local minima (marked by '$*$' in the figure) at $\lambda \approx 0.9$ for the E/H/1/K case and $\lambda \approx 1.1$ for the H/E/1/K case.

A further observation is that when $c_A^2 = c_S^2$, the BRAVO effect appears to have the same "magnitude" as that of the M/M/1/K case: a reduction of the asymptotic variance rate by a factor of $\frac{2}{3}$ for large K. This observation is demonstrated in Fig. 9 where we summarize results of several PH/PH/1/K systems with service and inter-arrival distributions having SCV: $\frac{1}{2}$, 1, $\frac{6}{5}$, $\frac{3}{2}$, 2. These are calculated using Erlang, exponential and hyper-exponential distributions as before.

6 Discussion

This paper was motivated by the practical question of calculating the asymptotic variance rate of the output of finite birth and death queues. We found that this variance rate is optimized when the input rate and service rate are balanced. In deriving these

results we discovered some unexpected phenomena, for which we do not yet have sufficient explanations.

Firstly, there is the "$\frac{2}{3}$ phenomenon", which we proved for M/M/1/K: in summary, when $\rho = 1$ and $K \to \infty$ the asymptotic variance rates of the outputs and of the overflows are the same and equal $\frac{2}{3}\lambda$ and this is possibly true for any choice of distribution of service and inter-arrival times as long as $c_A^2 = c_S^2$. We note that the value of $\frac{2}{3}$ for the asymptotic variance rate of the overflow process has been well known, as in the formula (22) which is due to Berger and Whitt. See also Theorem 5.7.4 of [35], as well as [37]. A well known fact is that the asymptotic variance of integrated Brownian motion with $\sigma^2 = 2$ is $\frac{2}{3}$ (cf. [25]). We do not see an immediate connection here but suspect there may be one.

Further surprises which our analytic and numeric results show took the form of singularities that occur in the M/M/1/K queue at the point $\rho = 1$ when $K \to \infty$: (a) The y-intercept of the linear asymptote of the variance function is maximized and approaches a delta function. (b) The limiting correlation coefficient between the outputs and the overflows exhibits a sharp change of sign. (c) The graph of $\overline{T}_\lambda$ has a singular point, dropping from the values of $\approx K + 1$ to 0. It is plausible that all these are closely related, and may hold for general inter-arrival and service distributions. The details are yet to be discovered.

Acknowledgements We thank Ward Whitt for pointing out his results that are summarized in Proposition 2.2. We also thank Itay Gurvich for some useful comments.

References

1. Asmussen, S.: Applied Probability and Queues. Springer, Berlin (2003)
2. Barnes, J.A., Disney, R.L.: Traffic processes in a class of finite Markovian queues. Queueing Syst. **6**, 311–326 (1990)
3. Berger, A.W., Whitt, W.: The Brownian approximation for rate-control throttles and the G/G/1/C queue. Discrete Event Dyn. Syst. Theory Appl. **2**, 7–60 (1992)
4. Branford, A.J.: On a property of finite-state birth and death processes. J. Appl. Probab. **23**, 859–866 (1986)
5. Breuer, L., Baum, D.: An Introduction to Queueing Theory and Matrix-Analytic Methods. Springer, Berlin (2005)
6. Chandramohan, J., Foley, R.D., Disney, R.L.: Thinning of point processes—covariance analysis. Adv. Appl. Probab. **17**, 127–146 (1985)
7. Cinlar, E., Disney, R.L.: Stream of overflows from a finite queue. Oper. Res. **15**(1), 131–134 (1967)
8. Cox, D.R., Isham, V.: Point Processes. Chapman and Hall, London (1980)
9. Daley, D.J.: Queueing output processes. Adv. Appl. Probab. **8**, 395–415 (1976)
10. Disney, R.L., de Morais, P.R.: Covariance properties for the departure process of M/Ek/1/N queues. AIIE Trans. **8**(2), 169–175 (1976)
11. Disney, R.L., Kiessler, P.C.: Traffic Processes in Queueing Networks—A Markov Renewal Approach. Johns Hopkins University Press, Baltimore (1987)
12. Disney, R.L., Konig, D.: Queueing networks: A survey of their random processes. SIAM Rev. **27**(3), 335–403 (1985)
13. Fischer, W., Meier-Hellstern, K.: The Markov-modulated Poisson process (MMPP) cookbook. Perform. Eval. **18**, 149–171 (1992)
14. Gershwin, S.B.: Variance of output of a tandem production system. In: Onvural, R., Akyildiz, I. (Eds.) Queueing Networks with Finite Capacity. Proceedings of the Second International Conference on Queueing Networks with Finite Capacity. Elsevier, Amsterdam (1993)
15. He, Q., Neuts, M.F.: Markov chains with marked transitions. Stoch. Process. Appl. **74**, 37–52 (1998)

16. Hendricks, K.B.: The output processes of serial production lines of exponential machines with finite buffers. Oper. Res. **40**(6), 1139–1147 (1992)
17. Hendricks, K.B., McClain, J.O.: The output processes of serial production lines of general machines with finite buffers. Manag. Sci. **39**(10), 1194–1201 (1993)
18. Keilson, J.: Markov Chain Models—Rarity and Exponentiality. Springer, Berlin (1979)
19. Kelly, F.: Reversibility and Stochastic Networks. Wiley, New York (1979)
20. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM, Philadelphia (1999)
21. Miltenburg, G.J.: Variance of the number of units produced on a transfer line with buffer inventories during a period of length T. Nav. Res. Logist. **34**, 811–822 (1987)
22. Naryana, S., Neuts, M.F.: The first two moment matrices of the counts for the Markovian arrival process. Stoch. Models **8**(3), 459–477 (1992)
23. Neuts, M.F., Li, J.: The input/output process of a queue. Appl. Stoch. Models Bus. Ind. **16**, 11–21 (2000)
24. Parthasarathy, P.R., Sudhesh, R.: The overflow process from a state-dependent queue. Int. J. Comput. Math. **82**(9), 1073–1093 (2005)
25. Parzen, E.: Stochastic Processes. Holden–Day, Oakland (1962)
26. Pourbabai, B.: Approximation of the overflow process from a G/M/N/K queueing system. Manag. Sci. **33**(7), 931–938 (1987)
27. Reynolds, J.F.: The covariance structure of queues and related processes—a survey of recent work. Adv. Appl. Probab. **7**, 383–415 (1975)
28. Rudemo, M.: Point processes generated by transitions of Markov chains. Adv. Appl. Probab. **5**, 262–286 (1973)
29. Tan, B.: Variance of the output as a function of time: Production line dynamics. Eur. J. Oper. Res. **177**(3), 470–484 (1999)
30. Tan, B.: Asymptotic variance rate of the output in production lines with finite buffers. Ann. Oper. Res. **93**, 385–403 (2000)
31. van Doorn, E.A.: On the overflow process from a finite Markovian queue. Perform. Eval. **4**, 233–240 (1984)
32. Whitt, W.: Approximating a point process by a renewal process, I: Two basic methods. Oper. Res. **30**(1), 125–147 (1982)
33. Whitt, W.: The queueing network analyzer. Bell Syst. Tech. J. **62**(9), 2779–2815 (1983)
34. Whitt, W.: Asymptotic formulas for Markov processes with applications to simulation. Oper. Res. **40**(2), 279–291 (1992)
35. Whitt, W.: Stochastic-Process Limits, an Introduction to Stochastic-Process Limits and their Application to Queues. Springer, Berlin (2001)
36. Whitt, W.: Heavy traffic limits for loss proportions in single-server queues. Queueing Syst. **46**, 507–536 (2004)
37. Williams, R.J.: Asymptotic variance parameters for the boundary local times of reflected Brownian motion on a compact interval. J. Appl. Probab. **29**(4), 996–1002 (1992)