mathematics

MDPI

Article

Analysis of Discrete-Time Queues with Branching Arrivals

Dieter Fiems *,† and **Koen De Turck** †

Department Telin, Ghent University, B-9000 Gent, Belgium
* Correspondence: dieter.fiems@ugent.be
† These authors contributed equally to this work.

Abstract: We consider a discrete-time single server queueing system, where arrivals stem from a multi-type Galton–Watson branching process with migration. This branching-type arrival process exhibits intricate correlation, and the performance of the corresponding queueing process can be assessed analytically. We find closed-form expressions for various moments of both the queue content and packet delay. Close inspection of the arrival process at hand, however, reveals that sample paths consist of large independent bursts of arrivals followed by geometrically distributed periods without arrivals. Allowing for non-geometric periods without arrivals, and correlated bursts, we apply π-thinning on the arrival process. As no closed-form expressions can be obtained for the performance of the corresponding queueing system, we focus on approximations of the main performance measures in the light and heavy traffic regimes.

Keywords: discrete-time queue; arrival correlation; branching process

MSC: 60K25

check for updates

Citation: Fiems, D.; De Turck, K. Analysis of Discrete-Time Queues with Branching Arrivals. *Mathematics* **2023**, *11*, 1020. https://doi.org/10.3390/math11041020

Academic Editor: Alexander Zeifman

Received: 13 January 2023
Revised: 10 February 2023
Accepted: 15 February 2023
Published: 16 February 2023

1. Introduction

Input traffic at intermediate routers in packet-switched communication networks typically exhibits considerable time correlation. Similar arrival correlation is also observed for arrival processes of goods or customers at queues in manufacturing and logistic systems. Correlated arrivals lead to increased variability in the arrival rate, which can make it more difficult to predict and manage the flow of goods and customers through the system. This results in longer queues, increased waiting times, and reduced system capacity. Moreover, correlated arrivals can also lead to increased variability in the utilization of resources, which in turn leads to increased maintenance costs and decreased efficiency.

Arrival correlation significantly affects queueing performance [1], and there is a continuing interest in analytically tractable queueing models with arrival correlation. Models of interest include, in particular, Markovian arrival models. Such models may have a finite state space, such as the discrete-time batch-Markovian arrival model [2–4], or a structured infinite state space, such as the discrete autoregressive arrival models [5,6] and the train and session arrival models [7,8]. For queues with an unstructured finite state space arrival model, the performance measures of interest are typically not available in closed form. Instead, numerically efficient algorithms are devised that yield the numerical values of the various performance measures once all model parameters are specified. In contrast, for some models with infinite state spaces, structural properties of the state space allow for expressions of the performance measures of interest in closed form.

In this paper, a Markovian arrival process is proposed with an infinite state space that is both structured and multi-dimensional: the Galton–Watson arrival process. More precisely, we assess the performance of a discrete-time queueing system, where the arrivals during the time slots are modeled as the size of the consecutive generations in multi-type Galton–Watson branching processes with migration [9–12]. Galton–Watson branching processes were originally investigated to study the extinction of family names, and constitute a

convenient abstraction for modeling the evolution of populations over time. Some recent applications of branching processes include pandemic spread [13], in vitro bactericidal potency [14], antibody affinity maturation [15], the dismantling of terrorist networks [16] and earthquake occurrence [17].

The main motivation for studying queues with Galton–Watson arrival processes stems from a combination of the analytical tractability of the queue content and packet delay and the observation that this arrival process exhibits intricate arrival correlation. The arrival process at hand generalizes train arrival models [7,8] with phase-type distributed train lengths. Moreover, there is a connection between branching processes and Hawkes processes, which can serve as arrival processes in infinite server queues [18,19]. Not unlike train and session arrival models, the Galton–Watson model is in a fixed state when there are no arrivals. It is, in fact, this property which enables closed-form expressions for the moments of the queue content and delay [20]. This also implies that periods without arrivals are geometrically distributed, and that the single-server queue is overloaded whenever there are arrivals. To overcome these rather stringent assumptions, we also investigate the queueing system with branching arrivals after applying π-thinning on the arrival process. With π-thinning, each packet arrival is discarded with a fixed probability so that the state of the arrival process is no longer fixed in the absence of arrivals. Unfortunately, no closed-form expressions can be obtained for the performance of the corresponding queueing system. Therefore, we focus on the light traffic and heavy traffic regimes. Our approach extends the results of [21], which investigated π-thinning of a queue with geometric train arrivals.

While Galton–Watson arrival processes are new, various queueing processes can be modeled as multi-type branching processes. First, the time till extinction in a branching process relates to busy periods in queues [22,23]. The branching property here follows from the observation that during a customer's service, new customers arrive that bring additional work. Secondly, many polling systems can be modeled as multi-type branching processes [24–26], the branching property of polling disciplines being a key facilitator of the analysis [27]. Finally, multi-type branching processes can also model the dynamics of infinite server queues [28].

The remainder of the paper is organized as follows. In the next section, the assumptions and notation of the arrival process are introduced, as well as the queueing model at hand. Following a probability generating the functions approach in Section 3, we find expressions for performance measures of interest, such as the moments of the queue content and the waiting times. Our results are then illustrated by some numerical examples in Section 4, where we also discuss a simple arrival correlation fitting procedure. We then consider the branching arrival model subject to π-thinning in the light traffic regime and heavy traffic regime in Sections 5 and 6, respectively. Finally, conclusions are drawn in Section 7.

2. Queueing Model

We consider a discrete-time queueing system; time is divided into fixed-length intervals or slots. Incoming packets are stored in an unlimited buffer and serviced in the order they arrived, with fixed service times equal to the slot length. Service is synchronized with respect to slot boundaries, so packets cannot be serviced during their arrival slot.

Packet arrivals stem from a sub-critical multi-type Galton–Watson process with migration; the number of (arriving packet) types of this process is denoted by K, and the number of arriving packets of type k during slot n is denoted by $X_n^{(k)}$. We can then express the number of arriving packets of type k during slot n in terms of the number of arrivals of the different types at the previous slot as follows:

$$X_n^{(k)} = \sum_{i=1}^{K} \sum_{j=1}^{X_{n-1}^{(i)}} M_{j,n}^{(i,k)} + N_n^{(k)} . \tag{1}$$

Here, $M_{j,n}^{(i,k)}$ is the type k off-spring of the jth type i packet at slot $n-1$, and $N_n^{(k)}$ denotes the number of new type k packets at slot n. The total number of packet arrivals A_n during slot n is then simply the sum of the arrivals of the different types:

$$A_n = \sum_{k=1}^{K} X_n^{(k)} . \tag{2}$$

In (1), the vectors $\left\{ \left[M_{j,n}^{(i,1)}, M_{j,n}^{(i,2)}, \dots M_{j,n}^{(i,K)} \right],\ j,n = 1,2,\dots \right\}$ constitute a doubly indexed sequence of independent and identically distributed (iid) random vectors for all $i \in \{1,2,\dots,K\}$. To characterize this sequence, we need to specify the joint distribution or the joint probability generating function of the offspring of the different types for an arrival of type $i \in \{1,2,\dots,K\}$. Hence, the random vectors are completely characterized by the vector-valued joint probability generating function:

$$\mathbf{M}(\mathbf{x}) = [M_i(\mathbf{x})]_{i=1}^{K} = \left[\mathsf{E}\left[\prod_{k=1}^{K} x_k^{M_{j,n}^{(i,k)}} \right] \right]_{i=1,2,\dots,K} , \tag{3}$$

with $\mathbf{x} = [x_1, x_2, \dots, x_K]$. Similarly, the vectors $\left\{ \left[N_n^{(1)}, N_n^{(2)}, \dots, N_n^{(K)} \right],\ n = 1,2,\dots \right\}$ constitute a sequence of iid random vectors, characterized by the common joint probability generating function:

$$N(\mathbf{x}) = \mathsf{E}\left[\prod_{k=1}^{K} x_k^{N_n^{(k)}} \right] . \tag{4}$$

Finally, we introduce notation for the mean offspring and migration of the different types. Let $\mu_{ik} = \mathsf{E}\left[M_{1,1}^{(i,k)} \right]$ denote the mean type k offspring of a type i packet. Similarly, $\nu_k = \mathsf{E}\left[N_1^{(k)} \right]$ denotes the mean number of new type k arrivals in a slot. For notational convenience, we collect these averages in a $K \times K$ matrix $\mathcal{M} = [\mu_{ik}]$ and in a column vector $\mathcal{V} = [\nu_k]$, respectively. With this notation, the mean number of arrivals in a slot equals

$$\rho = \mathsf{E}[A_1] = \mathbf{e}\,(\mathcal{I} - \mathcal{M})^{-1}\mathcal{V} . \tag{5}$$

Here, $\mathbf{e}$ denotes a row vector of ones and $\mathcal{I}$ is the identity matrix. As we assume that the multi-type branching process is sub-critical, the spectral radius of $\mathcal{M}$ is smaller than one. Hence, $(\mathcal{I} - \mathcal{M})^{-1}$ exists, and the mean arrival load is finite. This assumption will be retained in the following sections.

3. Queueing Analysis

With the notation of the arrival process established, we now focus on the queueing analysis. Let U_n denote the queue content at the beginning of slot n. The queue contents at the beginning of consecutive slots are then related as follows:

$$U_n = (U_{n-1} - 1)^+ + A_n . \tag{6}$$

Here, $(\cdot)^+$ is the usual shorthand notation for $\max(\cdot, 0)$.

The state of the queueing system at slot boundary n is completely described in the Markovian sense by the vector of state variables $(U_n, X_n^{(1)}, \dots, X_n^{(K)})$; see Equations (1), (2) and (6). Following a standard Loynes-type argument [29], the Markov process exhibits a unique stationary (and limiting) distribution for $\rho < 1$. Recall that (5) expresses ρ in terms of the parameters of the branching process with migration. Therefore, let $P(\mathbf{x}, z)$ denote the joint probability generating function of this vector in steady state:

$$P(\mathbf{x}, z) = \lim_{n\to\infty} \mathsf{E}\left[\prod_{k=1}^{K} x_k^{X_n^{(k)}} z^{U_n} \right] . \tag{7}$$

In view of Equations (1), (2) and (6) and by standard z-transform techniques, it is found that $P(\mathbf{x}, z)$ satisfies the following functional equation:

$$P(\mathbf{x}, z) = [P(\mathbf{M}(\mathbf{x}z), z) - P(\mathbf{M}(\mathbf{x}z), 0)]\frac{N(\mathbf{x}z)}{z} + P(\mathbf{M}(\mathbf{x}z), 0)N(\mathbf{x}z)\,.$$

Noting that $U_n = 0$ implies $A_n = 0$ and therefore also $X_n^{(k)} = 0$ for $k = 1, \ldots, K$, we have $P(\mathbf{x}, 0) = P(\mathbf{0}, 0)$, with $\mathbf{0}$ a row vector of zeros. The functional equation therefore simplifies to

$$P(\mathbf{x}, z) = P(\mathbf{M}(\mathbf{x}z), z)\frac{1}{z}N(\mathbf{x}z) - P(\mathbf{0}, 0)\frac{1-z}{z}N(\mathbf{x}z)\,. \tag{8}$$

The remaining unknown constant $P(\mathbf{0}, 0)$ follows from the normalization condition $P(\mathbf{e}, 1) = 1$. More precisely, multiplying both sides of Equation (8) by z and evaluating the first-order derivatives of the resulting expression in $(\mathbf{e}, 1)$ yields a system of linear equations for $P(\mathbf{0}, 0)$ and for the derivatives $\partial/\partial x_k P(\mathbf{x}, 1)|_{\mathbf{x}=\mathbf{e}}$. For $\rho < 1$, this system of equations has a unique solution. In particular, $P(\mathbf{0}, 0)$ equals

$$P(\mathbf{0}, 0) = 1 - \rho\,.$$

This is not entirely unexpected, as $P(\mathbf{0}, 0)$ represents the probability that the server is idle.

By recursive substitution, we can also obtain an explicit expression for the joint generating function $P(\mathbf{x}, z)$. To this end, recursively define the row vector $\mathbf{Q}^{(i)}(\mathbf{x}, z)$ as follows:

$$\mathbf{Q}^{(i)}(\mathbf{x}, z) = \mathbf{M}(\mathbf{Q}^{(i-1)}(\mathbf{x}, z)z)\,, \quad \mathbf{Q}^{(0)}(\mathbf{x}, z) = \mathbf{x}\,, \tag{9}$$

for $i = 1, 2, \ldots$. Successive application of the functional Equation (8) then yields

$$P(\mathbf{x}, z) = (1-\rho)(z-1)\sum_{j=0}^{\infty}\prod_{i=0}^{j}\frac{N(\mathbf{Q}^{(i)}(\mathbf{x}, z)z)}{z}\,. \tag{10}$$

Clearly, the probability generating function of the queue content equals $U(z) = P(\mathbf{e}, z)$. Additionally, the probability generating function for the queue content relates to the probability generating function for the packet delay. We can easily obtain the probability generating function for the packet delay (i.e., the number of slots between the arrival and departure of a packet) using the distributional form of Little's result for discrete-time queues with single-slot service times [30]:

$$D(z) = \frac{1}{\rho}(P(\mathbf{e}, z) - (1-\rho))\,. \tag{11}$$

Finally, exploiting the moment-generating property of probability generating functions immediately yields expressions for the various moments of the queue content and packet delay. Note that expressions of the derivatives of $P(\mathbf{x}, z)$ in $\mathbf{x} = \mathbf{e}$ and $z = 1$ can also be directly obtained by evaluating the derivatives of the functional Equation (8) in $\mathbf{x} = \mathbf{e}$ and $z = 1$. Compared to evaluating the derivatives of (10), such an approach is more convenient, as there is no need to evaluate the derivatives of the infinite sum.

4. Numerical Results

With the formulas at hand, we now investigate the mean delay of the queueing system with multi-type Galton–Watson arrivals. We focus on a simplified arrival model, where the arrivals stem from a two-type Galton–Watson model with neither migration between the types ($\mu_{12} = \mu_{21} = 0$) nor correlation between the new arrivals of the different types. For this arrival process, we introduce a convenient parameter estimation procedure.

The load ρ and the autocorrelation function $\alpha(n)$ of this arrival process are equal to

$$\rho = \frac{\nu_1}{1-\mu_{11}} + \frac{\nu_2}{1-\mu_{22}}, \quad \alpha(n) = \frac{\phi_1^2}{\phi^2}\,{\mu_{11}}^n + \frac{\phi_2^2}{\phi^2}\,{\mu_{22}}^n, \qquad \phi_i^2 = \frac{\theta_i^2(1-\mu_{ii}) + \nu_i\sigma_i^2}{(1+\mu_{ii})(1-\mu_{ii})^2}. \quad (12)$$

Here, ϕ_i^2 is the variance of the number of type-i arrivals in a slot, and $\phi^2 = \phi_1^2 + \phi_2^2$ is the variance of the number of arrivals in a slot. Further, σ_i^2 and θ_i^2 denote the variances of $M_{1,1}^{(i,i)}$ and $N_1^{(i)}$, respectively. To limit the number of parameters, assume that (i) $M_{1,1}^{(i,i)}$ is Bernoulli distributed such that $\sigma_i^2 = \mu_{ii}(1-\mu_{ii})$ and that (ii) $N_{1,1}^{(1)}$ and $N_{1,1}^{(2)}$ have the same index of dispersion (or variance-to-mean ratio), $\beta = \theta_1^2/\nu_1 = \theta_2^2/\nu_2$. Figure 1 displays the autocorrelation function of the arrival process for various parameter settings. The tangent $\alpha_0(n)$ in 0 and the asymptote $\alpha_\infty(n)$ are depicted as well in the logarithmic plot. These are given by

$$\begin{aligned} \alpha_0(n) &= \exp(n(\kappa \ln \mu_{11} + (1-\kappa)\ln \mu_{22})), \\ \alpha_\infty(n) &= \exp(n \ln \mu_{11} + \ln(\kappa)). \end{aligned} \quad (13)$$

with $\kappa = \phi_1^2/\phi^2$. Here, we assumed $\mu_{11} > \mu_{22}$, without loss of generality.

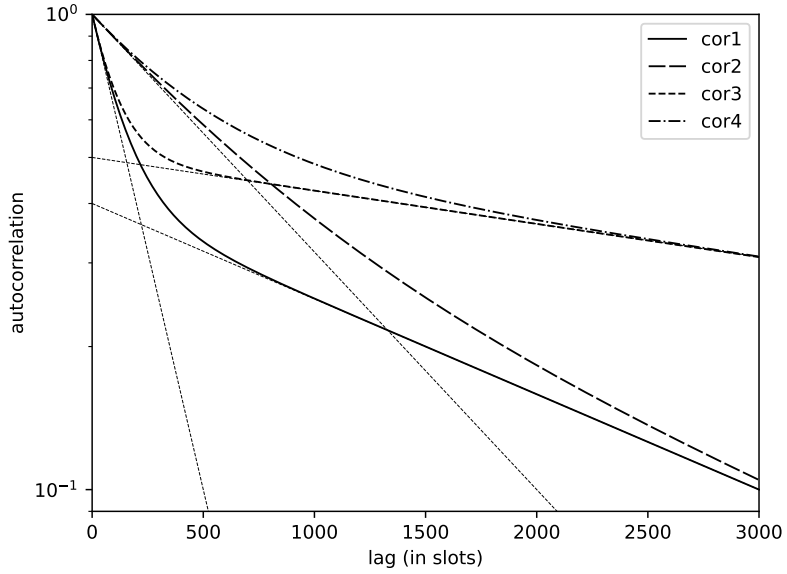

Figure 1. Autocorrelation function of the arrival process for different parameter settings: $(\kappa, \mu_{11}, \mu_{22})$ = (0.4, 0.999538, 0.992660) for cor1, (0.4, 0.999538, 0.998391) for cor2, for (0.5, 0.999839, 0.990991) cor3 and (0.5, 0.999839, 0.997861) for cor4.

The selection of the tangent and asymptote can be used to find suitable parameters for the arrival process. Modifying the decay rate of the tangent and the asymptote changes the long- and short-time correlation in the arrival process. The selection of the tangent and asymptote uniquely determines the parameters κ, μ_{11} and μ_{22}, which, in turn, uniquely determine the autocorrelation function; see Equations (12) and (13). It now suffices to additionally fix the total arrival load ρ and the variance of the number of arrivals in a slot ϕ^2. We can then express the remaining parameters ν_1, ν_2 and β in terms of ρ, ϕ^2, κ, μ_{11} and μ_{22}:

$$\nu_1 = (1-\mu_{11})(r+\rho), \quad \nu_2 = -(1-\mu_{22})r, \quad \beta = -\frac{1}{r}(1+\mu_{22})(1-\kappa)\phi^2 - \mu_{22},$$

where r is the unique root in $[-\rho, 0]$ of the quadratic equation:

$$(\mu_{11} - \mu_{22})r^2 - \Big[(1 + \kappa\mu_{11} + (1-\kappa)\mu_{22})\phi^2 - \rho(\mu_{11} - \mu_{22})\Big]r - \phi^2\rho(1-\mu_{22})(1-\kappa) = 0\,.$$

The solution is a valid parameter set for the process at hand, provided that $\beta \geq \max(1-\nu_1, 1-\nu_2)$. This is possible provided that ϕ^2 is sufficiently large, and a simple sufficient condition for that is $\phi^2 \geq \rho(1-\kappa)^{-1}$.

Summarizing, the procedure above can be used to match the characteristics of the multi-type branching arrival process with the corresponding characteristics of a traffic trace. That is, parameter estimation for a given traffic trace reduces to (i) the estimation of the mean and variance of the number of arrivals in a slot and to (ii) estimating $\alpha_0(n)$ and $\alpha_\infty(n)$ for the empirical autocorrelation function of the trace.

Figure 2 displays the mean delay as a function of the arrival load ρ, for an index of dispersion $\phi^2/\rho = 2$, and for the various autocorrelation curves from Figure 1. Recall that by using the estimation procedure described above, all arrival process parameters can be determined given the load, index of dispersion, and autocorrelation curve. Figure 2 illustrates the impact of the arrival correlation on the performance measures, highlighting the significant negative impact of a slow decay of the long-term correlation (cor3 and cor4) on system performance. Short-term correlation (cor2 and cor4) also results in performance degradation, although to a lesser extent than long-term correlation.

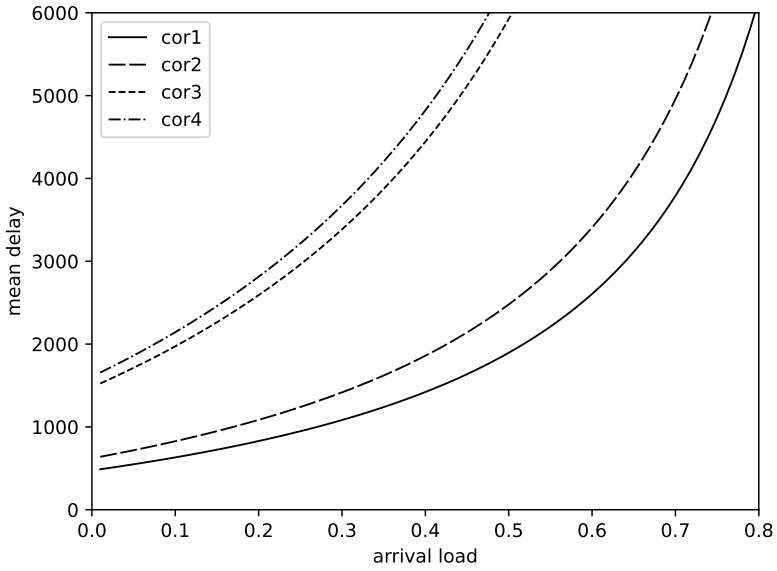

Figure 2. Mean packet delay vs. the arrival rate for the parameter settings of Figure 1 and for an index of dispersion $\phi^2/\rho = 2$.

5. Light-Traffic Analysis

In the analysis of Section 3, it is key that $A_n = 0$ implies $X_n^{(k)} = 0$ for $k = 1, \ldots, K$. More precisely, as the queue can only become empty when there are no new arrivals, the arrival process is always in state $\mathbf{0}$, when the queue becomes empty. Hence, we have $P(\mathbf{x}, 0) = P(\mathbf{0}, 0)$, and one only needs to determine a constant $P(\mathbf{0}, 0)$ instead of the unknown partial joint probability generating function $P(\mathbf{x}, 0)$.

While the modeling assumptions above allow for intricate correlation, the direct mapping of the population size on the number of arrivals does impose some limitations: whenever the total population size is larger than 1, the queue is overloaded, such that the arrival process will be in state $\mathbf{0}$ most of the time for $\rho < 1$ (which is needed for stability). Therefore, the present section considers the case where π-thinning is applied to the arrival process. That is, each individual of the population in a slot will contribute a packet to

the arrival process with probability q, and does not contribute to the arrival process with probability $1-q$. The number of arrivals A_n^q during slot n can therefore be expressed as

$$A_n^q = \sum_{\ell=1}^{A_n} Y_n^{(\ell)}\,, \tag{14}$$

where A_n is defined as before, see Equations (1) and (2), and where $\{Y_n^{(m)}\}$ is a double-indexed sequence of independent Bernoulli-distributed random variables with success probability q. As in Section 3, let U_n^q denote the queue content at the nth slot boundary, where the superscript makes the thinning probability q explicit. As a single packet departs in a slot, we again have

$$U_n^q = (U_{n-1}^q - 1)^+ + A_n^q\,. \tag{15}$$

As for the queueing model without thinning, the process $(X_n^{(1)}, \ldots, X_n^{(K)}, U_n^q)$ constitutes a Markov process. Let $P_q(\mathbf{x}, z)$ denote the probability generating function of the stationary distribution of the Markov process:

$$P_q(\mathbf{x}, z) = \lim_{n\to\infty} \mathsf{E}\left[\prod_{k=1}^{K} x_k^{X_n^{(k)}} z^{U_n^q}\right]. \tag{16}$$

Using standard z-transform techniques, we find the following functional equation for this generating function:

$$\begin{aligned} P_q(\mathbf{x}, z) = \frac{1}{z}\big(&P_q(\mathbf{M}(\mathbf{x}(1-q+qz)), z) \\ &+ (z-1)P_q(\mathbf{M}(\mathbf{x}(1-q+qz)), 0)\big)N(\mathbf{x}(1-q+qz))\,. \end{aligned} \tag{17}$$

In contrast to the model of Section 3, there is no straightforward solution for this functional equation. Therefore, we focus on the system's performance in the light-traffic regime. More precisely, we determine the terms in the series expansion of $P_q(\mathbf{x}, z)$ around $q = 0$. To this end, we consider the following series expansion of $P_q(\mathbf{x}, z)$:

$$P_q(\mathbf{x}, z) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{x})\, q^k z^\ell\,. \tag{18}$$

Note that there are no terms in $q^k z^\ell$ for $\ell > k$. This follows from the so-called n-events rule. The nth term in the series expansion of the stationary probability of a state is zero if this state cannot be reached by at most n q-events from a state which is reachable for the case $q = 0$. In our case, q-events are arrivals, and the queue content cannot grow beyond n with at most n arrivals.

As $P_q(\mathbf{x}, z)$ is a probability function for each q, the normalization condition $P_q(\mathbf{e}, 1) = 1$ leads to the following normalization conditions for the terms $T_{k,\ell}$:

$$T_{0,0}(\mathbf{e}) = 1\,, \quad \sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{e}) = 0\,. \tag{19}$$

We now show that we can recursively retrieve all derivatives of the terms $T_{k,\ell}$ in $\mathbf{z} = \mathbf{e}$. Substituting the series expansion (18) in the functional Equation (17) yields

$$\begin{aligned} \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{x}) q^k z^{\ell+1} = &\sum_{k=0}^{\infty}\sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz)))q^k z^\ell N(\mathbf{x}(1-q+qz)) \\ &+ (z-1)\sum_{k=0}^{\infty} T_{k,0}(\mathbf{M}(\mathbf{x}(1-q+qz)))q^k N(\mathbf{x}(1-q+qz))\,. \end{aligned} \tag{20}$$

In order to compare terms in equal powers of q at both sides of the equation, we need to introduce the Maclaurin series expansion of the terms $T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz)))$ and $N(\mathbf{x}(1-q+qz))$. For the latter, the expansion reads

$$N(\mathbf{x}(1-q+qz)) = \sum_{n=0}^{\infty} q^n (z-1)^n \widehat{N}^{(n)}(\mathbf{x}),$$

with,

$$\widehat{N}^{(n)}(\mathbf{x}) = \sum_{|\mathbf{m}|=n} \frac{1}{\mathbf{m}!} N^{(\mathbf{m})}(\mathbf{x}) \mathbf{x}^{\mathbf{m}}.$$

In the expression above, the sum runs over all nth-order derivatives of N, and $N^{(\mathbf{m})}(\mathbf{x})$, $\mathbf{x}^{\mathbf{m}}$, and $\mathbf{m}!$ are shorthands for,

$$N^{(\mathbf{m})}(\mathbf{x}) = \frac{\partial^{m_1}}{\partial x_1^{m_1}} \cdots \frac{\partial^{m_K}}{\partial x_K^{m_K}} N(\mathbf{x}), \quad \mathbf{x}^{\mathbf{m}} = \prod_{k=1}^{K} x_k^{m_k}, \quad \mathbf{m}! = \prod_{k=1}^{K} m_k!,$$

for $\mathbf{m} = [m_1, \ldots, m_K]$. The expansion of $T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz)))$ is somewhat more involved. First note, that

$$\frac{d^n}{dq^n} M_i(\mathbf{x}(1-q+qz))\Big|_{q=0} = (z-1)^n \widehat{M}_i^{(n)}(\mathbf{x}),$$

with,

$$\widehat{M}_i^{(n)}(\mathbf{x}) = \sum_{|\mathbf{m}|=n} \frac{n!}{\mathbf{m}!} M_i^{(\mathbf{m})}(\mathbf{x}) \mathbf{x}^{\mathbf{m}}.$$

Moreover, let $\widehat{\mathbf{M}}^{(n)}(\mathbf{x})$ be the vector with entries $\widehat{M}_i^{(n)}(\mathbf{x})$, $i = 1, \ldots, K$. We can then use the multivariate extension of Faa di Bruno's formula [31] to calculate the nth derivative of q in $q = 0$:

$$\frac{d^n}{dq^n} T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz)))\Big|_{q=0} = (z-1)^n \widehat{T}_{k,\ell}^{(n)}(\mathbf{x}),$$

with,

$$\widehat{T}_{k,\ell}^{(n)}(\mathbf{x}) = \sum_{1 \le |\mathbf{m}| \le n} T_{k,\ell}^{(\mathbf{m})}(\mathbf{x}) \sum_{\theta(n,\mathbf{m})} n! \prod_{j=1}^{n} \frac{\left(\widehat{\mathbf{M}}^{(j)}(\mathbf{x})\right)^{\mathbf{k}_j}}{\mathbf{k}_j! (j!)^{|\mathbf{k}_j|}},$$

and,

$$\theta(n,\mathbf{m}) = \left\{ (\mathbf{k}_1, \ldots, \mathbf{k}_n) : \mathbf{k}_j \in \mathbb{N}^K, \sum_{j=1}^{n} \mathbf{k}_j = \mathbf{m}, \sum_{j=1}^{n} j|\mathbf{k}_j| = n \right\}.$$

Finally, the Maclaurin series expansion reads

$$T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz))) = \sum_{n=0}^{\infty} \frac{1}{n!} q^n (z-1)^n \widehat{T}_{k,\ell}^{(n)}(\mathbf{x}).$$

Substituting the expansions of $T_{k,\ell}(\mathbf{M}(\mathbf{x}(1-q+qz)))$ and $N(\mathbf{x}(1-q+qz))$ in (20) yields

$$\begin{aligned} \sum_{k=0}^{\infty} \sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{x}) q^k z^{\ell+1} = {} & \sum_{k=0}^{\infty} \sum_{\ell=0}^{k} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{n!} \widehat{T}_{k,\ell}^{(n)}(\mathbf{x}) (z-1)^{m+n} z^{\ell} \widehat{N}^{(m)}(\mathbf{x}) q^{k+m+n} \\ & + \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{n!} \widehat{T}_{k,0}^{(n)}(\mathbf{x}) (z-1)^{m+n+1} \widehat{N}^{(m)}(\mathbf{x}) q^{k+m+n}. \end{aligned} \quad (21)$$

The former expression now allows for identifying the coefficients of the terms $q^{\kappa} z^{\lambda}$, yielding

$$T_{0,0}(\mathbf{x}) = T_{0,0}(\mathbf{M}(\mathbf{x})) N(\mathbf{x}),$$

for $\kappa = \lambda = 0$ and,

$$T_{\kappa,\lambda-1}(\mathbf{x}) = T_{\kappa,\lambda}(\mathbf{M}(\mathbf{x}))N(\mathbf{x}) + T_{\kappa,0}(\mathbf{M}(\mathbf{x}))N(\mathbf{x})1_{\{\lambda=1\}} - T_{\kappa,0}(\mathbf{M}(\mathbf{x}))N(\mathbf{x})1_{\{\lambda=0\}} + F_{\kappa,\lambda}(\mathbf{x}) \quad (22)$$

with

$$F_{\kappa,\lambda}(\mathbf{x}) = \sum_{k=0}^{\kappa-1}\sum_{\ell=0}^{k}\sum_{n=0}^{\kappa-k}\frac{1}{n!}\widehat{T}_{k,\ell}^{(n)}(\mathbf{x})\sum_{m=0}^{\kappa-k}\binom{\kappa-k}{m}(-1)^{\kappa-k-m}\widehat{N}^{(\kappa-k-n)}(\mathbf{x})1_{\{\ell+m=\lambda\}} + \sum_{k=0}^{\kappa-1}\sum_{n=0}^{\kappa-k}\frac{1}{n!}\widehat{T}_{k,0}^{(n)}(\mathbf{x})\sum_{m=0}^{\kappa-k+1}\binom{\kappa-k+1}{m}(-1)^{\kappa-k+1-m}\widehat{N}^{(\kappa-k-n)}(\mathbf{x})1_{\{\lambda=m\}}, \quad (23)$$

for $\kappa > 0$ and $\lambda \leq \kappa$. Close inspection of the expression above shows that $F_{\kappa,\lambda}(\mathbf{x})$ only depends on $T_{k,\ell}$ for $k < \kappa$. Moreover, as $\mathbf{M}(\mathbf{e}) = \mathbf{e}$, we can retrieve all derivatives of $T_{\kappa,\ell}$ in $\mathbf{x} = \mathbf{e}$ from the system of functional Equation (22) and the normalization condition (19) in terms of the derivatives of $T_{k,\ell}$ for $k < \kappa$. Summarizing, we can recursively determine all derivatives $T_{k,\ell}^{(\mathbf{m})}(\mathbf{e})$.

Once these terms have been found, it is now trivial to calculate the light-traffic approximation of the mean queue content. From the series expansion (18), we immediately find the following expression for the Kth-order expansion of the nth moment of the queue content:

$$\mathsf{E}[U^n] \approx \sum_{k=0}^{K}\sum_{\ell=0}^{k} T_{k,\ell}(\mathbf{e})q^k\ell^n.$$

Finally, note that, as for the system without π-thinning, the distributional form of Little's result applies. Hence, we can express moments of the delay in terms of moments of the queue content, and our approximation of the moments of the queue content can be used to approximate the moments of the delay.

To illustrate the light-traffic analysis, we reconsider the numerical example of Section 4, albeit with different parameters to ensure that the system is in overload, prior to π-thinning. More specifically, we set $\rho = 2$, $\mu_{11} = 0.99$ and $\mu_{22} = 0.9$. It then suffices to specify the variance ϕ^2 of the number of arrivals in a slot and the thinning probability q to fully specify the arrival model. Figure 3 depicts the 7th-order light-traffic approximation of the mean delay as a function of the traffic load. The load is varied by varying the thinning probability q for fixed $\rho = 2$. To verify the accuracy of the light-traffic approximation, we also simulate the queueing system. The markers on the plot show the values of the mean delay as obtained by simulation. We rely on the deletion–replication approach (p. 241, [32]) with 100 replications of 10^8 slots to assess the accuracy of the simulation results but omit the confidence intervals from the plot, as the upper and lower boundaries of the confidence intervals are visually indistinguishable. The first 1000 slots of each run are deleted to reduce bias, in line with the method of Welch (p. 238, [32]).

Clearly, the light traffic approach is accurate for low loads, while the approximation deviates from the simulation results for higher loads. Additionally, more accurate results are obtained if the arrival variance is smaller. Additional experimentation with higher-order approximations shows that the approximation for low load further improves by increasing the order of the approximation, but this is not necessarily the case for higher load.

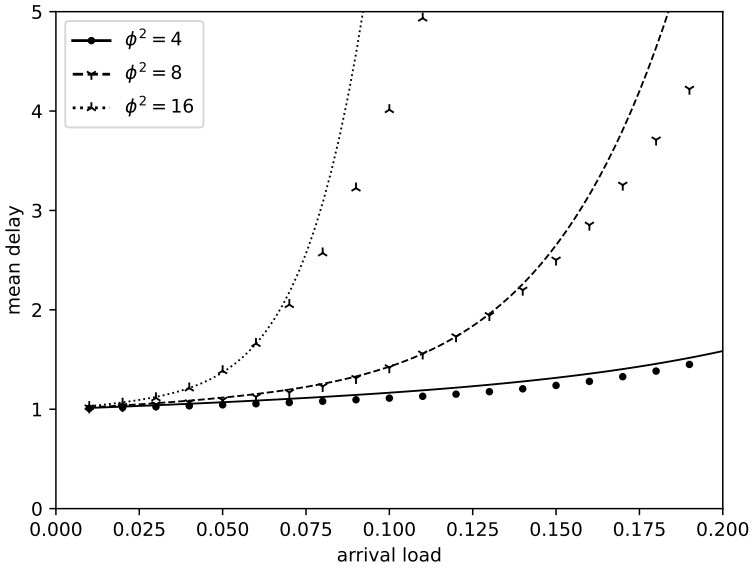

Figure 3. Light-traffic approximation of the mean delay as a function of the load for different values of the arrival variance ϕ^2 as indicated. The load is varied by varying the thinning probability.

6. Heavy Traffic Approximation

In line of the arguments of the preceding section, we again apply π-thinning on the Galton–Watson arrival process. As before, let q denote the probability that an individual in the population contributes a packet. While the preceding section considered the light-traffic regime, we now consider the heavy-traffic regime as pioneered by Kingman in the seminal paper [33]. To this end, we assume that the arrival rate prior to thinning exceeds 1, $\rho > 1$. In this case, the queue is not stable for $q = 1$, and we study the limit $q \to q_m = \rho^{-1}$.

Following the exposition in chapter 9 of [34], the heavy traffic performance measures are given in terms of the load ρ and the asymptotic variance V:

$$V = \lim_{k\to\infty} \frac{1}{k} \mathsf{Var}\left[\sum_{n=1}^{k} A_n^{q_m}\right],$$

where $\{A_n^{q_m}\}$ is the stationary arrival process with π-thinning with parameter q_m. By the stationarity of the arrival process, the former expression simplifies to

$$V = \mathsf{Var}[A_1^{q_m}] + \lim_{k\to\infty} \frac{2}{k} \sum_{m=1}^{k} \sum_{n=1}^{m-1} \mathsf{Cov}[A_{m-n+1}^{q_m}, A_1^{q_m}].$$

We now express the variance V in terms of the process $\{A_n\}$ without π-thinning. In view of the definition of $A_n^{q_m}$, see Equation (14), we have

$$V = \mathsf{Var}\left[\sum_{\ell=1}^{A_1} Y_1^{(\ell)}\right] + \lim_{k\to\infty} \frac{2}{k} \sum_{m=1}^{k} \sum_{n=1}^{m-1} \mathsf{Cov}\left[\sum_{\ell=1}^{A_{m=1+1}} Y_{m-n+1}^{(\ell)}, \sum_{\ell'=1}^{A_1} Y_1^{(\ell')}\right],$$

where Y_n^{ℓ} is an independent Bernoulli random variable with success probability q_m for each index n and ℓ. By conditioning on the unthinned arrival process $\{A_n\}$ and by noting that $\{Y_n^{(\ell)}\}$ constitutes a doubly indexed sequence of independent and identically distributed random variables, the variance and covariances in the former expression simplify to

$$V = \mathsf{Var}[A_1]\mathsf{E}[Y_1^{(1)}]^2 + \mathsf{E}[A_1]\mathsf{Var}[Y_1^{(1)}] + \lim_{k\to\infty} \frac{2\mathsf{E}[Y_{m-n+1}^{(1)}]\mathsf{E}[Y_1^{(1)}]}{k} \sum_{m=1}^{k} \sum_{n=1}^{m-1} \mathsf{Cov}[A_{m-n+1}, A_1].$$

As $Y_n^{(\ell)}$ is a Bernoulli random variable with success probability q_m for each index n and ℓ, we have $\mathsf{E}[Y_n^{(\ell)}] = q_m$ and $\mathsf{Var}[Y_n^{(\ell)}] = q_m(1-q_m)$ such that

$$V = \mathsf{Var}[A_1]q_m^2 + \mathsf{E}[A_1]q_m(1-q_m) + \lim_{k\to\infty} \frac{2q_m^2}{k} \sum_{m=1}^{k} \sum_{n=1}^{m-1} \mathsf{Cov}[A_{m-n+1}, A_1]\,. \tag{24}$$

By conditioning on the number of arrivals of the different types in the consecutive slots, we have

$$\mathsf{Cov}[A_{m-n+1}, A_1] = \mathbf{e}\mathcal{M}^{m-n}\mathcal{C}\mathbf{e}'\,, \tag{25}$$

where $\mathcal{M} = [\mu_{ik}]$ is the matrix of the mean offspring of the different types as introduced in Section 2 and where $\mathbf{e}'$ is a column vector of ones. Moreover, $\mathcal{C} = [c_{ik}]$ is the covariance matrix of the branching arrival process. We can easily express the covariance in terms of the joint probability generating function $P(\mathbf{x}, z)$:

$$c_{ik} = \frac{\partial^2}{\partial x_i \partial x_k} P(\mathbf{x}, 1)\bigg|_{\mathbf{x}=\mathbf{e}} - \frac{\partial}{\partial x_i} P(\mathbf{x}, 1)\bigg|_{\mathbf{x}=\mathbf{e}} \frac{\partial}{\partial x_k} P(\mathbf{x}, 1)\bigg|_{\mathbf{x}=\mathbf{e}}\,.$$

Plugging the expression for the covariance (25) in (24), we finally find

$$V = \mathbf{e}\,\mathcal{C}\,\mathbf{e}'q_m^2 + \rho q_m(1-q_m) + 2q_m^2\mathbf{e}\,\mathcal{M}(\mathcal{I}-\mathcal{M})^{-1}\mathcal{C}\,\mathbf{e}'\,.$$

Having determined the asymptotic variance, we can now have the following expressions for the first two moments of the buffer content under heavy traffic:

$$\mathsf{E}[U_{HT}] = \frac{V}{2(1-q\rho)}\,, \quad \mathsf{E}[U_{HT}^2] = \frac{V^2}{2(1-q\rho)^2}\,.$$

To evaluate the accuracy of the heavy-traffic limit, we reconsider the numerical example of the light-traffic approximation: we set $\rho = 2$, $\mu_{11} = 0.99$ and $\mu_{22} = 0.9$. Table 1 lists the values of the scaled queue content $\mathsf{E}[U(1-\rho q)]$ for different values of the load ρq up to 98%, as well as the corresponding heavy traffic limit value $V/2$. The pre-limit values are obtained by simulation, using a replication–deletion approach with 100 replications, each replication simulating 10^8 slots, where the first 1000 slots of each run are discarded to reduce the bias. Table 1 also lists the 99% confidence interval. The results show that the scaled mean queue content indeed converges to the heavy traffic limit.

Table 1. The scaled queue content $\mathsf{E}[U(1-\rho q)]$ for different values of the load ρq, and the corresponding heavy traffic limit value $V/2$.

	$\rho q = 95\%$	$\rho q = 96\%$	$\rho q = 97\%$	$\rho q = 98\%$	$\rho q = 99\%$	$V/2$
$\phi^2 = 4$	39.5 ± 0.3	40.6 ± 0.4	42.4 ± 0.5	42.8 ± 0.9	45.3 ± 1.4	45.75
$\phi^2 = 8$	80.6 ± 0.9	81.6 ± 1.0	85.3 ± 1.6	85.8 ± 2.0	88.7 ± 4.5	91.25
$\phi^2 = 16$	161.8 ± 2.8	164.7 ± 3.0	169.1 ± 4.0	174.5 ± 6.2	170.9 ± 13.18	182.25

7. Conclusions

This paper presented closed-form expressions for the probability generating functions of the queue content and packet delay in a discrete-time queueing system, where arrivals stem from a multi-type Galton–Watson branching process. The model is useful for analyzing buffers with significant arrival correlation, which was demonstrated by a parameter estimation procedure for a simplified Galton–Watson-type arrival process. We showed that the parameters of this simplified Galton–Watson-type arrival process for a given empirical auto-correlation function are easily estimated. Moreover, once these parameters are known, evaluating the relevant performance measures is straightforward. We then applied π-thinning on the Galton–Watson arrival process. In contrast to the unthinned process,

closed-form expressions for the various performance measures of interest are no longer available. We therefore studied the corresponding thinned queueing system in both the light-traffic and heavy-traffic regimes. Finally, to verify the accuracy of the approximations in these regimes, we simulated the queueing process with thinning.

Author Contributions: Conceptualization, D.F. and K.D.T.; methodology, D.F. and K.D.T.; software, D.F.; validation, D.F. and K.D.T.; formal analysis, D.F. and K.D.T.; writing—original draft preparation, D.F. and K.D.T.; writing—review and editing, D.F. and K.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schwefel, H.P.; Antonios, I.; Lipsky, L. Understanding the relationship between network traffic correlation and queueing behavior: A review based on the N-Burst ON/OFF model. *Perform. Eval.* **2017**, *115*, 68–91. [CrossRef]
2. Blondia, C.; Casals, O. Statistical multiplexing of VBR sources: A matrix-analytic approach. *Perform. Eval.* **1992**, *16*, 5–20. [CrossRef]
3. Herrmann, C. The complete analysis of the discrete time finite DBMAP/G/1/N queue. *Perform. Eval.* **2001**, *43*, 95–121. [CrossRef]
4. Pradhan, S.; Gupta, U. Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. *Ann. Oper. Res.* **2019**, *277*, 161–196. [CrossRef]
5. Hwang, G.U.; Sohraby, K. On the exact analysis of a discrete-time queueing system with autoregressive inputs. *Queueing Syst.* **2003**, *43*, 29–41. [CrossRef]
6. Kamoun, F. The discrete-time queue with autoregressive inputs revisited. *Queueing Syst.* **2006**, *54*, 185–192. [CrossRef]
7. Wittevrongel, S. Discrete-time buffers with variable-length train arrivals. *Electron. Lett.* **1998**, *34*, 1719–1721. [CrossRef]
8. Hoflack, L.; De Vuyst, S.; Wittevrongel, S.; Bruneel, H. Analytic traffic model of web server. *Electron. Lett.* **2008**, *44*, 61–62. [CrossRef]
9. Athreya, K.; Ney, P. *Branching Processes*; Springer: Berlin/Heidelberg, Germany, 1972.
10. Kevei, P.; Wiandt, P. Moments of the stationary distribution of subcritical multitype Galton-Watson processes with immigration. *Stat. Probab. Lett.* **2021**, *173*, 109067. [CrossRef]
11. Barczy, M.; Nedenyi, F.; Pap, G. On aggregation of multitype Galton-Watson branching processes with immigration. *Mod. Stoch. Theory Appl.* **2018**, *5*, 53–79. [CrossRef]
12. Dyakonova, E.E. Multitype Galton-Watson branching processes in Markovian random environment. *Theory Probab. Its Appl.* **2012**, *56*, 508–517. [CrossRef]
13. Minzer, D.; Oz, Y.; Safra, M.; Wainstain, L. Pandemic spread in communities via random graphs. *J. Stat.-Mech.-Theory Exp.* **2021**, *2021*, 113501. [CrossRef]
14. Bogdanov, A.; Kevei, P.; Szalai, M.; Virok, D. Stochastic Modeling of In Vitro Bactericidal Potency. *Bull. Math. Biol.* **2022**, *84*, 6. [CrossRef] [PubMed]
15. Balelli, I.; Milisic, V.; Wainrib, G. Multi-type Galton-Watson Processes with Affinity-Dependent Selection Applied to Antibody Affinity Maturation. *Bull. Math. Biol.* **2019**, *81*, 830–868. [CrossRef]
16. Collins, B.; Hoang, D.T.; Nguyen, N.T.; Hwang, D. A New Model for Predicting and Dismantling a Complex Terrorist Network. *IEEE Access* **2022**, *10*, 126466–126478. [CrossRef]
17. Kovchegov, Y.; Zaliapin, I.; Ben-Zion, Y. Invariant Galton-Watson branching process for earthquake occurrence. *Geophys. J. Int.* **2022**, *231*, 567–583. [CrossRef]
18. Koops, D.T.; Saxena, M.; Boxma, O.J.; Mandjes, M. Infinite-server queues with Hawkes input. *J. Appl. Probab.* **2018**, *55*, 920–943. [CrossRef]
19. Selvamuthu, D.; Tardelli, P. Infinite-server systems with Hawkes arrivals and Hawkes services. *Queueing Syst.* **2022**, *101*, 329–351. [CrossRef]
20. Fiems, D.; De Turck, K. The Mean Queue Content of Discrete-time Queues with Zero-regenerative Arrivals. *Oper. Res. Lett.* **2012**, *40*, 235–238. [CrossRef]
21. Fiems, D.; De Turck, K. Taylor-series approximations for queues with arrival correlation. *Appl. Math. Model.* **2019**, *69*, 113–126. [CrossRef]
22. Nakayama, M.K.; Shahabuddin, P.; Sigman, K. On finite exponential moments for branching processes and busy periods for queues. *J. Appl. Probab.* **2004**, *41A*, 273–280. [CrossRef]
23. Ernst, P.A.; Asmussen, S.; Hasenbein, J.J. Stability and busy periods in a multiclass queue with state-dependent arrival rates. *Queueing Syst.* **2018**, *90*, 207–224. [CrossRef]
24. Altman, E.; Fiems, D. Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Syst.* **2007**, *56*, 241–253. [CrossRef]

25. Vatutin, V.A. Polling systems and multitype branching processes in random environment with final product. *Theory Probab. Its Appl.* **2011**, *55*, 631–660. [CrossRef]
26. Fiems, D.; Altman, E. Gated polling with stationary ergodic walking times, Markovian routing and random feedback. *Ann. Oper. Res.* **2012**, *198*, 145–164. [CrossRef]
27. Resing, J. Polling systems and multi-type branching processes. *Queueing Syst.* **1993**, *13*, 409–426. [CrossRef]
28. Altman, E. On stochastic recursive equations and infinite server queues. In Proceedings of the IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, FL, USA, 13–17 March 2005.
29. Baccelli, F.; Bremaud, P. *Elements of Queueing Theory*; Springer: Berlin/Heidelberg, Germany, 1994.
30. Vinck, B.; Bruneel, H. Delay analysis for single server queues. *Electron. Lett.* **1996**, *32*, 802–803. [CrossRef]
31. Constantine, G.; Savits, T. A multivariate Faa di Bruno formula with applications. *Trans. Am. Math. Soc.* **1996**, *348*, 503–520. [CrossRef]
32. Banks, J. (Ed.) *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*; John Wiley & Sons: Hoboken, NJ, USA, 1998.
33. Kingman, J. On queues in heavy traffic. *J. R. Stat. Soc. Ser. B (Methodol.)* **1962**, *24*, 383–392. [CrossRef]
34. Whitt, W. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*; Springer Series in Operations Research and Financial Engineering; Springer: New York, NY, USA, 2011.