mathematics

MDPI

Article

Delay in a 2-State Discrete-Time Queue with Stochastic State-Period Lengths and State-Dependent Server Availability and Arrivals

Freek Verdonck *, Herwig Bruneel and Sabine Wittevrongel

SMACS Research Group, Department of Telecommunications and Information Processing (TELIN), Ghent University (UGent), Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium; herwig.bruneel@ugent.be (H.B.); sabine.wittevrongel@ugent.be (S.W.)

* Correspondence: freek.verdonck@ugent.be

Abstract: In this paper, we consider a discrete-time multiserver queueing system with correlation in the arrival process and in the server availability. Specifically, we are interested in the delay characteristics. The system is assumed to be in one of two different system states, and each state is characterized by its own distributions for the number of arrivals and the number of available servers in a slot. Within a state, these numbers are independent and identically distributed random variables. State changes can only occur at slot boundaries and mark the beginnings and ends of state periods. Each state has its own distribution for its period lengths, expressed in the number of slots. The stochastic process that describes the state changes introduces correlation to the system, e.g., long periods with low arrival intensity can be alternated by short periods with high arrival intensity. Using probability generating functions and the theory of the dominant singularity, we find the tail probabilities of the delay.

Keywords: queueing theory; discrete-time; multiserver; correlation; delay; tail

Citation: Verdonck, F.; Bruneel, H.; Wittevrongel, S. Delay in a 2-State Discrete-Time Queue with Stochastic State-Period Lengths and State-Dependent Server Availability and Arrivals. *Mathematics* **2021**, *9*, 1709. https://doi.org/10.3390/math9141709

Academic Editor: Konstantin Samouylov

Received: 17 June 2021
Accepted: 19 July 2021
Published: 20 July 2021

1. Introduction

When, in the early 20th century, the Danish mathematician Agner Erlang used a mathematical model to describe a telephone switch (at the time, an office where workers manually connected phone lines), he became the founder of the field of queueing theory. More than a hundred years later, queueing theory is used in a broad range of practical problems such as in traffic engineering and industrial engineering, but telecommunications remains one of the main fields of application [1–3].

Multiserver queueing models have received considerable attention in the past. Most papers, however, focus on the analysis of the queue length characteristics only. Results are available for queueing systems with a constant number of available servers [4–6] as well as for queueing systems with a variable number of available servers [7–9].

When it comes to delay analysis, multiserver queueing systems are notoriously hard to analyze, but some results do exist in the literature. In [10–12], discrete-time queueing systems are treated where the number of available servers is constant, while in [13], the number of available servers in a slot is a stochastic variable which is independent and identically distributed (i.i.d.) from slot to slot. In [14], a continuous-time queueing system is treated where N servers are subject to breakdowns and repair; the results are the distributions of the queue length and waiting times. The authors of [15] obtain the queue length distribution and the delay distribution for a discrete-time model with general service demands and correlated service capacities. In their model, the service time of a customer depends on both its service demand and the service capacity.

In this paper, we consider a discrete-time multiserver queueing system that is special in the way that it allows correlation in both the slot-to-slot server availability and the arrival

process. The buffer capacity is assumed to be infinite and the queueing discipline is First Come First Served (FCFS). In our earlier paper [16], we handled the system content of such a queueing system, and now we focus on the analysis of the delay characteristics. Table 1 compares the setting and delay results of the current paper to those of the relevant related papers indicated above.

Table 1. Brief summary of the setting and delay results of relevant papers in comparison with the current paper.

Paper	Servers	Arrivals	Delay Results
Chaudhry et al. [10]	constant	general (bulk), i.i.d.	full delay distribution
Bruneel et al. [11]	constant	general, i.i.d.	tail probabilities
Gao et al. [12]	constant	general, i.i.d.	tail probabilities
Laevens et al. [13]	i.i.d.	general, i.i.d.	tail probabilities
Neuts et al. [14]	i.i.d.	general, i.i.d.	full delay distribution
De Muynck et al. [15]	correlated	general, i.i.d.	full delay distribution
The current paper	correlated	correlated	tail probabilites

The considered queueing system can be in two different system states. For every slot, a stochastic number of servers is available, and a stochastic number of new customers arrive. These are both i.i.d. within each state, but these distributions can be different for each state. The system resides for a stochastic number of slots in the same state before state transitions occur, which can only happen at slot boundaries. The sojourn times follow a distribution which is also dependent on the system state.

The above setup can be used to model a wide variety of queueing systems, such as queues with bursty arrivals (where long periods of low/no arrival intensity are alternated by short periods of high arrival intensity) and queues with cyclical service capacities (where alternately few and many servers are available for fixed period lengths).

The potential applications can be found in many fields. In [17,18], continuous time models are used to calculate the delay in a system with time-varying arrival intensity and service capacity. The envisaged application is a hospital emergency department, but the model can also be used in other applications. In [19], time-dependent arrival rates are called a key feature of many real life service systems and it is, therefore, included in the Erlang loss model which is presented. In [20], an overview of time-varying queueing models is presented, from a telecommunications point of view. A possible application of the method of this paper is a processor sharing system with speed scaling [21], where the processing speed is adapted to the (expected) load. Another possible application is the domain of mobile ad hoc networks [22]. In such networks, nodes (which can move freely) must cooperate to send and receive packages. The number of available nodes as well as the traffic varies over time.

The outline of the paper is as follows. In Section 2, we give the mathematical outline of the queueing model. In Section 3, we repeat some key results regarding the system content from our earlier paper [16]. In Section 4, we condition the delay of a customer on the state of the arrival slot and on the number of customers present in the queue at the moment of arrival. Section 5 then describes how the delay of an arbitrary customer can be obtained. In Section 6, we use the theory of the dominant singularity to obtain the tail characteristics of the delay. The numerical examples in Section 7 illustrate the model and Section 8 concludes the paper.

2. Queueing Model

We consider the discrete-time multiserver queueing model of [16] with correlation over time in both the arrival process and the server availability. In this discrete-time model, the time horizon is divided into slots of equal length and, during a slot, the system can be either in state 1 or state 2. Note that, in general, in our paper, we will always use i to

indicate a system state and not repeat that it can only take values 1 and 2. Furthermore, we will use $\bar{\imath}$ to refer to the *other* state:

$$\bar{\imath} \triangleq 3 - i\,. \tag{1}$$

Therefore, put simply, there are two types of slots, which we call state-1-slots and state-2-slots, and during a state-i-slot the queueing system is in system state i. State changes can only occur at slot boundaries and mark the beginnings and ends of state-1-periods and state-2-periods. If we denote with $r_{i,k}$ the length (expressed in the number of slots) of the kth state-i-period, then the series $\{r_{i,k}\}$ form two sets of i.i.d. stochastic variables. Their distribution is given by

$$r_i(n) \triangleq \text{Prob}[\text{state-}i\text{-period has } n \text{ slots}]\,, \quad n \geq 1\,; \tag{2}$$

$$R_i(z) \triangleq \sum_{n=1}^{\infty} r_i(n) z^n\,; \quad \overline{r}_i \triangleq \sum_{n=1}^{\infty} n r_i(n) = R_i'(1)\,, \tag{3}$$

where we have introduced the probability generating functions (pgfs) $R_i(z)$. We limit the $R_i(z)$ to be rational functions of their argument:

$$R_i(z) = \frac{A_r^i(z)}{B_r^i(z)}\,, \tag{4}$$

with $A_r^i(z)$ and $B_r^i(z)$ mutually prime polynomials of degree m_{Ar}^i and m_{Br}^i, respectively, and with $A_r^i(1) = B_r^i(1) = 1$. We define m_r^i as:

$$m_r^i \triangleq \max(m_{Ar}^i, m_{Br}^i)\,. \tag{5}$$

The probability that an arbitrary slot belongs to a given state is equal to the fraction of time the system is in that state and is given by

$$\sigma_i \triangleq \text{Prob}[\text{arbitrary slot belongs to state } i] = \frac{\overline{r}_i}{\overline{r}_1 + \overline{r}_2}\,. \tag{6}$$

The special feature of our model is that the server availability and the arrival process both depend on the system state. Specifically, the distribution of the number of available servers during a slot depends on the system state during that slot in the following way:

$$s_i(n) \triangleq \text{Prob}[n \text{ servers available during a state-}i\text{-slot}]\,, \quad n \geq 1\,; \tag{7}$$

$$S_i(z) \triangleq \sum_{n=1}^{\infty} s_i(n) z^n\,; \quad \overline{s}_i \triangleq \sum_{n=1}^{\infty} n s_i(n) = S_i'(1)\,. \tag{8}$$

During every slot, there is at least one server available and within a given state-i-period the numbers of available servers during the different slots are i.i.d. from slot to slot. Similarly to the $R_i(z)$, we also limit the $S_i(z)$ to be rational functions of their argument:

$$S_i(z) = \frac{A_s^i(z)}{B_s^i(z)}\,, \tag{9}$$

with $A_s^i(z)$ and $B_s^i(z)$ mutually prime polynomials of degree m_{As}^i and m_{Bs}^i, respectively, and with $A_s^i(1) = B_s^i(1) = 1$. We define m_s^i as:

$$m_s^i \triangleq \max(m_{As}^i, m_{Bs}^i)\,. \tag{10}$$

The numbers of arrivals during the different slots of a given state-i-period are i.i.d. from slot to slot as well. Their common distribution is characterized by

$$c_i(n) \triangleq \text{Prob}[\text{state-}i\text{-slot has } n \text{ arrivals}], \quad n \geq 0; \tag{11}$$

$$C_i(z) \triangleq \sum_{n=0}^{\infty} c_i(n) z^n; \quad \lambda_i \triangleq \sum_{n=1}^{\infty} n c_i(n) = C_i'(1). \tag{12}$$

The average arrival intensity is then given by

$$\lambda \triangleq \sigma_1 \lambda_1 + \sigma_2 \lambda_2. \tag{13}$$

Customers can only start service at slot boundaries, so an arriving customer can only be taken into service at the beginning of the next slot, even if a server is idle at the moment of arrival. The queue capacity is assumed to be infinite, so an arriving customer will always join the system. Each customer requires exactly one slot of service.

We assume the system reaches steady state and, therefore, the average number of customers entering the system should be strictly smaller than the average number of available servers [23], leading to the following stability condition:

$$\lambda < \sigma_1 \bar{s}_1 + \sigma_2 \bar{s}_2. \tag{14}$$

Furthermore, it will prove useful to introduce the following notation:

$$Y_i(z) \triangleq C_i(z) S_i\left(\frac{1}{z}\right). \tag{15}$$

3. System Content

In an earlier work [16], we analyzed the system content for a queueing model as described above. In the current section, we repeat the main results that are necessary for the delay analysis in this paper. Let us denote with the stochastic variable g_k^i ($k \geq 0$), the total number of customers in the system at the beginning of the $(k+1)$st slot of a state-i-period. The corresponding pgf is $G_k^i(z)$. We can derive the following recursive equation, valid for $k \geq 1$:

$$G_k^i(z) = Y_i(z) G_{k-1}^i(z) + C_i(z) \sum_{l=0}^{\infty} \sum_{j=1}^{\infty} \text{Prob}[g_{k-1}^i = l] s_i(l+j) \left(1 - z^{-j}\right). \tag{16}$$

We can obtain a set of two linear equations for the functions $G_0^i(z)$ by recursive application of (16) and by stating that the system content at the beginning of a state-i-period equals the system content at the end of a state-$\bar{\imath}$-period. This set of equations can then be solved to yield the following expression:

$$G_0^i(z) = \frac{S_{\bar{\imath}}\left(\frac{1}{z}\right) R_{\bar{\imath}}(Y_{\bar{\imath}}(z)) \left[Q_i(Y_i(z), 1) - Q_i\left(Y_i(z), \frac{1}{z}\right)\right] + S_i\left(\frac{1}{z}\right) \left[Q_{\bar{\imath}}(Y_{\bar{\imath}}(z), 1) - Q_{\bar{\imath}}\left(Y_{\bar{\imath}}(z), \frac{1}{z}\right)\right]}{S_i\left(\frac{1}{z}\right) S_{\bar{\imath}}\left(\frac{1}{z}\right) [1 - R_i(Y_i(z)) R_{\bar{\imath}}(Y_{\bar{\imath}}(z))]}, \tag{17}$$

with the bivariate functions $Q_i(x,z)$ unknown. It can be proven that if $R_i(z)$ and $S_i(z)$ are rational functions of their argument, then also the $Q_i(x,z)$ are rational and of the following form:

$$Q_i(x,z) = \frac{A_q^i(x,z)}{B_r^i(x) B_s^i(z)} \quad ; \quad A_q^i(x,z) \triangleq \sum_{n=1}^{m_r^i} \sum_{j=1}^{m_s^i} \epsilon_{nj}^i x^n z^j. \tag{18}$$

The (finite number of) unknowns ϵ_{nj}^i can be determined by relying on the properties of pgfs, namely that they are normalized and that they are analytical within the complex unit

disk. To do this, the roots within the complex unit disk of the denominator of (17) need to be determined and a set of linear equations needs to be solved. For many common choices of the distributions in the model, this does not require large computational effort. Similarly, it is possible to obtain the pgfs of the system content at the beginning of an arbitrary state-i-slot and at the beginning of an arbitrary slot. Based on these pgfs, important performance metrics can be calculated, such as the expected number of customers in the system.

4. Delay of a Customer with K Customers Ahead

For the system content, the specific queueing discipline is not of importance, as long as it is work conserving. However, for the delay analysis it needs to be specified. We will assume a First Come First Served (FCFS) policy. We do not specify the exact arrival instant of a customer within a slot and, therefore, define its delay as the time interval from the first slot boundary after the customer's arrival until the end of its service slot. This definition is illustrated in Figure 1. The delay thus consists of an integer number of slots and is at least one slot long. This setup is also referred to as a Late Arrival System with Delayed Access (LAS-DA) [24].

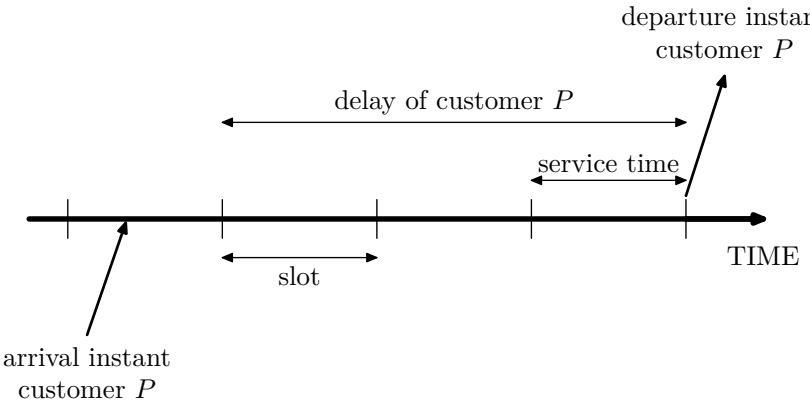

Figure 1. The delay of customer P.

Let us denote with the stochastic variable $d_k^{i,n}$ the delay of a customer that arrives during a state-i-slot, with n more slots until the next state-$\bar{\imath}$-slot and with k customers waiting in the queue with priority over the arriving customer (excluding the customer(s) currently in service, if any). The corresponding pgf is $D_k^{i,n}(z)$. Clearly, if during the slot after the considered customer's arrival slot more than k servers are available, the customer's delay will consist of exactly one slot. Otherwise, if during the first delay slot only $l \leq k$ servers are available, there will be an additional number of delay slots that corresponds to the delay of a customer with $k - l$ customers ahead. Based on these observations, we can state

$$D_0^{i,n}(z) = z\,; \tag{19}$$

$$D_k^{i,n}(z) = z \sum_{l=k+1}^{\infty} s_i(l) + z \sum_{l=1}^{k} D_{k-l}^{i,n-1}(z) s_i(l)\,, \quad k, n \geq 1\,; \tag{20}$$

$$D_k^{i,0}(z) = z \sum_{l=k+1}^{\infty} s_{\bar{\imath}}(l) + z \sum_{n=1}^{\infty} r_{\bar{\imath}}(n) \sum_{l=1}^{k} D_{k-l}^{\bar{\imath},n-1}(z) s_{\bar{\imath}}(l)\,, \quad k \geq 1\,. \tag{21}$$

Furthermore, we use the stochastic variable d_k^i for the delay of a customer arriving during the first slot of a state-i-period and with k customers in the queue with priority over the arriving customer. The corresponding pgf $D_k^i(z)$ can be expressed as

$$D_k^i(z) = \sum_{n=1}^{\infty} r_i(n) D_k^{i,n-1}(z)\,. \tag{22}$$

Let us now introduce two auxiliary functions

$$D^{i,n}(x,z) \triangleq \sum_{k=0}^{\infty} D_k^{i,n}(z)x^k\,; \tag{23}$$

$$\begin{aligned} D^{i}(x,z) &\triangleq \sum_{k=0}^{\infty} D_k^{i}(z)x^k \\ &= \sum_{n=1}^{\infty}\sum_{k=0}^{\infty} r_i(n) D_k^{i,n-1}(z)x^k = \sum_{n=1}^{\infty} r_i(n) D^{i,n-1}(x,z)\,. \end{aligned} \tag{24}$$

Using the Relations (19) and (20), we obtain

$$D^{i,n}(x,z) = zS_i(x)D^{i,n-1}(x,z) + z\frac{1-S_i(x)}{1-x}\,, \quad n \geq 1\,. \tag{25}$$

After recursive application this leads to

$$D^{i,n}(x,z) = [zS_i(x)]^n D^{i,0}(x,z) + z\frac{1-[zS_i(x)]^n}{1-zS_i(x)}\,\frac{1-S_i(x)}{1-x}\,, \quad n \geq 0\,. \tag{26}$$

Multiplying (26) with $r_i(n+1)$ and summing over all $n \geq 0$ we get

$$D^{i}(x,z) = \frac{1}{zS_i(x)} R_i(zS_i(x)) D^{i,0}(x,z) + z\frac{1-S_i(x)}{(1-x)[1-zS_i(x)]}\left[1 - \frac{R_i(zS_i(x))}{zS_i(x)}\right]. \tag{27}$$

Using (21) we can also work out an expression for $D^{i,0}(x,z)$:

$$D^{i,0}(x,z) = zS_{\bar{i}}(x)D^{\bar{i}}(x,z) + z\frac{1-S_{\bar{i}}(x)}{1-x}\,. \tag{28}$$

Combining (27) and (28) we then obtain the following set of equations:

$$D^{i}(x,z) = \frac{1}{S_i(x)} R_i(zS_i(x))\left[S_{\bar{i}}(x)D^{\bar{i}}(x,z) + \frac{1-S_{\bar{i}}(x)}{1-x}\right] + z\frac{1-S_i(x)}{(1-x)[1-zS_i(x)]}\left[1 - \frac{R_i(zS_i(x))}{zS_i(x)}\right], \tag{29}$$

which can be solved to find an explicit expression for $D^{i}(x,z)$:

$$D^{i}(x,z) = \frac{\widetilde{f}^{i}(x,z)}{\widetilde{g}^{i}(x,z)}\,, \tag{30}$$

with

$$\begin{aligned} \widetilde{f}^{i}(x,z) =& z^2 S_i(x)S_{\bar{i}}(x)R_i(zS_i(x))R_{\bar{i}}(zS_{\bar{i}}(x))[1-S_i(x)] + zS_i(x)^2 R_i(zS_i(x))R_{\bar{i}}(zS_{\bar{i}}(x)) \\ &+ z^2 S_i(x)^2 S_{\bar{i}}(x) - zS_{\bar{i}}(x)R_i(zS_i(x))R_{\bar{i}}(zS_{\bar{i}}(x)) - z^2 S_i(x)S_{\bar{i}}(x) \\ &+ R_i(zS_i(x))[R_{\bar{i}}(zS_{\bar{i}}(x)) + z - 1][S_{\bar{i}}(x) - S_i(x)] - zS_i(x)^2 + zS_i(x)\,; \end{aligned} \tag{31}$$

$$\widetilde{g}^{i}(x,z) = S_i(x)(1-x)[1-zS_i(x)][1-zS_{\bar{i}}(x)][1-R_i(zS_i(x))R_{\bar{i}}(zS_{\bar{i}}(x))]\,. \tag{32}$$

Note that $\widetilde{f}^{i}(x,z)$ is divisible by $S_i(x)$ as $R_i(0) = 0$. Bearing in mind that all pgfs are normalized, we can furthermore verify that $\widetilde{f}^{i}(x,z)$ is divisible by $(1-x)$, $(1-zS_i(x))$ and $(1-zS_{\bar{i}}(x))$. In order to rewrite (30) as a rational function of x, we divide numerator and denominator by those common factors. To remove the remaining poles in x in $\widetilde{g}^{i}(x,z)$, we multiply with the auxiliary function $u(x,z)$, which in view of (4) and (9) is defined as

$$u(x,z) \triangleq B_r^i(zS_i(x))B_r^{\bar{i}}(zS_{\bar{i}}(x))B_s^i(x)^{m_r^i}B_s^{\bar{i}}(x)^{m_r^{\bar{i}}}\,. \tag{33}$$

Therefore, we obtain

$$D^i(x,z) = \frac{f^i(x,z)}{g(x,z)}\,, \tag{34}$$

with

$$f^i(x,z) = \frac{\widetilde{f}^i(x,z)u(x,z)}{S_i(x)(1-x)[1-zS_i(x)][1-zS_{\bar{i}}(x)]}\,; \tag{35}$$

$$g(x,z) = [1-R_i(zS_i(x))R_{\bar{i}}(zS_{\bar{i}}(x))]u(x,z)\,. \tag{36}$$

The new denominator $g(x,z)$ of $D^i(x,z)$ is a polynomial in x of degree $M \triangleq m_s^i m_r^i + m_s^{\bar{i}} m_r^{\bar{i}}$ and the numerator is of degree $M-1$. We can do a partial fraction expansion of $D^i(x,z)$ based on its poles x_ϕ in x, which we assume distinct. Note that the x_ϕ ($\phi = 1,\ldots,M$) are functions of z, but for notational simplicity the argument is omitted. We can then write $D^i(x,z)$ as

$$D^i(x,z) = \sum_{\phi=1}^{M} \frac{f^i(x_\phi,z)}{g_x(x_\phi,z)(x-x_\phi)}\,, \tag{37}$$

with

$$g_x(x,z) \triangleq \frac{\partial}{\partial x} g(x,z)\,. \tag{38}$$

The above allows us to easily obtain an expression for $D_k^i(z)$, which is the pgf of the delay of a customer arriving during the first slot of a state-i-period with k customers waiting in the queue with priority over the considered customer:

$$D_k^i(z) = \frac{1}{k!}\frac{\partial^k}{\partial x^k} D^i(x,z)\bigg|_{x=0} = \sum_{\phi=1}^{M} \frac{-f^i(x_\phi,z)}{x_\phi^{k+1} g_x(x_\phi,z)}\,. \tag{39}$$

5. Delay of an Arbitrary Customer

Let us now consider the arbitrary customer P, arriving during slot S. We denote the probability that S is a state-i-slot with h^i and that it is the lth slot of a state-i-period of in total $l+n$ slots as $h_{l|n}^i$:

$$h^i \triangleq \text{Prob}[P \text{ arrives during state-}i\text{-slot}] = \frac{\lambda_i \overline{r}_i}{\lambda_1 \overline{r}_1 + \lambda_2 \overline{r}_2}\,; \tag{40}$$

$$\begin{aligned} h_{l|n}^i &\triangleq \text{Prob}[P \text{ arrives during } l\text{th slot of state-}i\text{-period with } n+l \text{ slots in total}] \\ &= h^i \frac{r_i(n+l)}{\overline{r}_i}\,. \end{aligned} \tag{41}$$

The pgf $F_i(z)$ of the number of customers that arrive during slot S but before customer P and given that S is a state-i-slot is known to be given by (see, e.g., [1])

$$F_i(z) = \frac{C_i(z)-1}{\lambda_i(z-1)}\,. \tag{42}$$

The queue content as experienced by P upon arrival, are the customers with priority over P that can start service after S. It thus consists of the customers that were present in the system at the beginning of S, minus those that receive service during S and plus those that arrived during S but before P. It is a stochastic variable that depends on the state of S and on the time since the last state change. Given that S is the lth slot of a state-i-period, we

will denote this stochastic variable as t_l^i ($l \geq 1$). Its pgf $T_l^i(z)$ can be obtained by translating the above observations into the z-domain:

$$\begin{aligned} T_l^i(z) &= F_i(z) \sum_{k=0}^{\infty} \text{Prob}\Big[g_{l-1}^i = k\Big] \left\{ \sum_{j=1}^{k} s_i(j) z^{k-j} + \sum_{j=1}^{\infty} s_i(k+j) \right\} \\ &= F_i(z) G_{l-1}^i(z) S_i\left(\frac{1}{z}\right) + F_i(z) \sum_{k=0}^{\infty} \sum_{j=1}^{\infty} \text{Prob}\Big[g_{l-1}^i = k\Big] s_i(k+j)(1-z^{-j}) \,. \end{aligned} \tag{43}$$

We will denote the inverse z-transform of the above pgf as $t_l^i(k)$. We can now develop the pgf $W(z)$ of the delay of an arbitrary customer

$$W(z) = \sum_i h^i \sum_{n=0}^{\infty} \sum_{l=1}^{\infty} \frac{r_i(n+l)}{\bar{r}_i} \sum_{k=0}^{\infty} t_l^i(k) D_k^{i,n}(z) \,. \tag{44}$$

The functions $R_i(z)$ are assumed to be rational, if we further assume that they only have poles of multiplicity 1, we can write them in the form

$$R_i(z) = \sum_{j=1}^{M_i} \eta_j^i z^j + \sum_{j=1}^{N_i} \omega_j^i \frac{(1-\alpha_j^i)z}{1-\alpha_j^i z} \,. \tag{45}$$

Note that the summations in the above formula do not necessarily both appear. In the remainder of this paper, we will assume that both summations are present, the results can be easily modified for the other cases. The corresponding probability mass function (pmf) $r_i(n)$ can be written as

$$r_i(n) = \begin{cases} \eta_n^i + \sum_{j=1}^{N_i} \omega_j^i (1-\alpha_j^i) \left(\alpha_j^i\right)^{n-1}, & \text{if} \quad n \leq M_i \,; \\ \sum_{j=1}^{N_i} \omega_j^i (1-\alpha_j^i) \left(\alpha_j^i\right)^{n-1}, & \text{if} \quad n > M_i \,. \end{cases} \tag{46}$$

We substitute (46) into (44) and rearrange the summations to get:

$$W(z) = \sum_i \frac{h^i}{\bar{r}_i} \left\{ \sum_{j=1}^{M_i} \eta_j^i \sum_{l=1}^{j} \sum_{k=0}^{\infty} t_l^i(k) D_k^{i,j-l}(z) + \sum_{j=1}^{N_i} \omega_j^i \sum_{n=0}^{\infty} \sum_{l=1}^{\infty} (1-\alpha_j^i) \left(\alpha_j^i\right)^{n+l-1} \sum_{k=0}^{\infty} t_l^i(k) D_k^{i,n}(z) \right\}. \tag{47}$$

We will look at the above expression in more detail in two steps and, therefore, introduce the following auxiliary notations:

$$u_i^{j,l}(z) \triangleq \sum_{k=0}^{\infty} t_l^i(k) D_k^{i,j-l}(z) \,; \tag{48}$$

$$v_i^j(z) \triangleq \sum_{n=0}^{\infty} \sum_{l=1}^{\infty} \sum_{k=0}^{\infty} (1-\alpha_j^i) \left(\alpha_j^i\right)^{n+l-1} t_l^i(k) D_k^{i,n}(z) \,. \tag{49}$$

Let us first look at (48) for $l = j$. In that case, the arrival slot of customer P is the last slot of a state-i-period. We denote with n the number of servers available in the next slot (which is the first slot of a state-$\bar{\imath}$-period). With probability $t_j^i(k)$ there are k customers waiting in the queue with priority over P. The delay of the tagged customer P is 1 slot if $n > k$, or the pgf of its delay is given by $zD_{k-n}^{\bar{\imath}}(z)$ if $n \leq k$. In the z-domain, this yields

$$
\begin{aligned}
u_i^{j,j}(z) &= \sum_{n=1}^{\infty} s_{\bar{\imath}}(n) \sum_{k=0}^{n-1} t_j^i(k) z + \sum_{n=1}^{\infty} s_{\bar{\imath}}(n) \sum_{k=n}^{\infty} t_j^i(k) z D_{k-n}^{\bar{\imath}}(z) \\
&= \sum_{n=1}^{\infty} s_{\bar{\imath}}(n) \sum_{k=0}^{n-1} t_j^i(k) \left[z - \sum_{\phi=1}^{M} \frac{-z f^{\bar{\imath}}(x_\phi, z)}{x_\phi^{k-n+1} g_x(x_\phi, z)} \right] + \sum_{n=1}^{\infty} s_{\bar{\imath}}(n) \sum_{k=0}^{\infty} t_j^i(k) \sum_{\phi=1}^{M} \frac{-z f^{\bar{\imath}}(x_\phi, z)}{x_\phi^{k-n+1} g_x(x_\phi, z)} \\
&= \sum_{n=1}^{\infty} s_{\bar{\imath}}(n) \sum_{k=0}^{n-1} t_j^i(k) \left[z - \sum_{\phi=1}^{M} \frac{-z f^{\bar{\imath}}(x_\phi, z)}{x_\phi^{k-n+1} g_x(x_\phi, z)} \right] + \sum_{\phi=1}^{M} S_{\bar{\imath}}(x_\phi) T_j^i\left(\frac{1}{x_\phi}\right) \frac{-z f^{\bar{\imath}}(x_\phi, z)}{x_\phi g_x(x_\phi, z)}, \qquad (50)
\end{aligned}
$$

where we have also introduced (39). Similarly, we now look at the situation where $l = j - 1$, with $j > 1$. The arrival slot of customer P is the penultimate slot of a state-i-period. We denote with n_1 and n_2 the number of servers available in the next two slots (of which one is the last slot of a state-i-period and one is the first slot of a state-$\bar{\imath}$-period). With probability $t_{j-1}^i(k)$, there are k customers waiting in the queue with priority over the tagged customer P. The delay of P equals 1 slot if $n_1 > k$, the delay equals 2 slots if $n_1 \leq k < n_1 + n_2$ and the pgf of its delay is given by $z^2 D_{k-n_1-n_2}^{\bar{\imath}}(z)$ if $n_1 + n_2 \leq k$. In the z-domain, this yields

$$
\begin{aligned}
u_i^{j,j-1}(z) &= \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{k=0}^{n_1-1} t_{j-1}^i(k) z + \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{n_2=1}^{\infty} s_{\bar{\imath}}(n_2) \sum_{k=n_1}^{n_1+n_2-1} t_{j-1}^i(k) z^2 \\
&\quad + \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{n_2=1}^{\infty} s_{\bar{\imath}}(n_2) \sum_{k=n_1+n_2}^{\infty} t_{j-1}^i(k) z^2 D_{k-n_1-n_2}^{\bar{\imath}}(z) \\
&= \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{k=0}^{n_1-1} t_{j-1}^i(k) \left[z - \sum_{\phi=1}^{M} \frac{-z^2 f^{\bar{\imath}}(x_\phi, z) S_{\bar{\imath}}(x_\phi)}{x_\phi^{k-n_1+1} g_x(x_\phi, z)} \right] \\
&\quad + \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{n_2=1}^{\infty} s_{\bar{\imath}}(n_2) \sum_{k=n_1}^{n_1+n_2-1} t_{j-1}^i(k) \left[z^2 - \sum_{\phi=1}^{M} \frac{-z^2 f^{\bar{\imath}}(x_\phi, z)}{x_\phi^{k-n_1-n_2+1} g_x(x_\phi, z)} \right] \\
&\quad + \sum_{\phi=1}^{M} S_i(x_\phi) S_{\bar{\imath}}(x_\phi) T_{j-1}^i\left(\frac{1}{x_\phi}\right) \frac{-z^2 f^{\bar{\imath}}(x_\phi, z)}{x_\phi g_x(x_\phi, z)}. \qquad (51)
\end{aligned}
$$

The same reasoning can be applied to obtain an expression for the general function $u_i^{j,l}(z)$. There are $(j - l)$ full slots until the next state-$\bar{\imath}$-period. In these slots, $n_1, n_2, \ldots, n_{j-l}$ servers are available, with probabilities $s_i(n_1), s_i(n_2), \ldots, s_i(n_{j-l})$ and in the slot afterwards (the first slot of a state-$\bar{\imath}$-period), there are n_{j-l+1} servers available with probability $s_{\bar{\imath}}(n_{j-l+1})$. With probability $t_l^i(k)$ there are k customers waiting in the queue with priority over the tagged customer P. The delay of P equals s slots (with $1 \leq s \leq j - l + 1$) if $\sum_{p=1}^{s-1} n_p \leq k < \sum_{p=1}^{s} n_p$ and its delay is defined by the pgf $z^{j-l+1} D_{k-n_1-\ldots-n_{j-l+1}}^{\bar{\imath}}(z)$ if $\sum_{p=1}^{j-l+1} n_p \leq k$. In the z-domain, this yields

$$
\begin{aligned}
u_i^{j,l}(z) &= \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{k=0}^{n_1-1} t_l^i(k) \left[z - \sum_{\phi=1}^{M} \frac{-z^{j-l+1} f^{\bar{\imath}}(x_\phi, z) S_{\bar{\imath}}(x_\phi) S_i(x_\phi)^{j-l-1}}{x_\phi^{k-n_1+1} g_x(x_\phi, z)} \right] \\
&\quad + \sum_{n_1=1}^{\infty} s_i(n_1) \sum_{n_2=1}^{\infty} s_i(n_2) \sum_{k=n_1}^{n_1+n_2-1} t_l^i(k) \left[z^2 - \sum_{\phi=1}^{M} \frac{-z^{j-l+1} f^{\bar{\imath}}(x_\phi, z) S_{\bar{\imath}}(x_\phi) S_i(x_\phi)^{j-l-2}}{x_\phi^{k-n_1-n_2+1} g_x(x_\phi, z)} \right] \\
&\quad + \ldots \\
&\quad + \sum_{n_1=1}^{\infty} s_i(n_1) \ldots \sum_{n_{j-l}=1}^{\infty} s_i(n_{j-l}) \sum_{n_{j-l+1}=1}^{\infty} s_{\bar{\imath}}(n_{j-l+1}) \sum_{k=n_1+\ldots+n_{j-l}}^{n_1+\ldots+n_{j-l+1}-1} t_l^i(k) \left[z^{j-l+1} - \sum_{\phi=1}^{M} \frac{-z^{j-l+1} f^{\bar{\imath}}(x_\phi, z)}{x_\phi^{k-n_1-\ldots-n_{j-l+1}+1} g_x(x_\phi, z)} \right] \\
&\quad + \sum_{\phi=1}^{M} S_i(x_\phi)^{j-l} S_{\bar{\imath}}(x_\phi) T_l^i\left(\frac{1}{x_\phi}\right) \frac{-z^{j-l+1} f^{\bar{\imath}}(x_\phi, z)}{x_\phi g_x(x_\phi, z)}. \qquad (52)
\end{aligned}
$$

Now, we look at the second part of (47). In order to work out $v_i^j(z)$ as given in (49), we first introduce the following auxiliary functions:

$$D_{k,j}^i(z) \triangleq \sum_{n=0}^{\infty} (1-\alpha_j^i)\left(\alpha_j^i\right)^n D_k^{i,n}(z)\,; \tag{53}$$

$$D_j^i(x,z) \triangleq \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} (1-\alpha_j^i)\left(\alpha_j^i\right)^n D_k^{i,n}(z)x^k\,. \tag{54}$$

Taking first the sum over k in (54), we identify the definition of $D^{i,n}(x,z)$ as given in (23), so we can bring expression (26) into (54) and then work out the summation over n:

$$\begin{aligned} D_j^i(x,z) &= \sum_{n=0}^{\infty} (1-\alpha_j^i)\left(\alpha_j^i\right)^n \left\{ [zS_i(x)]^n D^{i,0}(x,z) + z\frac{1-[zS_i(x)]^n}{1-zS_i(x)}\frac{1-S_i(x)}{1-x}\right\} \\ &= \frac{1-\alpha_j^i}{1-\alpha_j^i zS_i(x)} D^{i,0}(x,z) + z\frac{1-S_i(x)}{(1-x)[1-zS_i(x)]}\left[1-\frac{1-\alpha_j^i}{1-\alpha_j^i zS_i(x)}\right]. \end{aligned} \tag{55}$$

Using the expression (28) for $D^{i,0}(x,z)$, we obtain

$$D_j^i(x,z) = \frac{1-\alpha_j^i}{1-\alpha_j^i zS_i(x)}\left[zS_{\bar{i}}(x)D^{\bar{i}}(x,z) + z\frac{1-S_{\bar{i}}(x)}{1-x}\right] + \alpha_j^i z\frac{1-S_i(x)}{(1-x)\left[1-\alpha_j^i zS_i(x)\right]}\,. \tag{56}$$

We then fill in the expression (34) for $D^{\bar{i}}(x,z)$ and rearrange to get

$$D_j^i(x,z) = \frac{f_j^i(x,z)}{g(x,z)}\,, \tag{57}$$

with $g(x,z)$ a polynomial in x of degree M as given in (36) and with

$$f_j^i(x,z) = \frac{(1-\alpha_j^i)z(1-x)S_{\bar{i}}(x)f^{\bar{i}}(x,z) + z\left[1-(1-\alpha_j^i)S_{\bar{i}}(x)-\alpha_j^i S_i(x)\right]g(x,z)}{(1-x)\left[1-\alpha_j^i zS_i(x)\right]}\,, \tag{58}$$

which can be shown to be a polynomial function in x of degree $M-1$. We can do a partial fraction expansion of $D_j^i(x,z)$ based on its poles x_ϕ in x. We can then obtain an expression for $D_{k,j}^i(z)$ as

$$D_{k,j}^i(z) = \sum_{\phi=1}^{M} \frac{-f_j^i(x_\phi,z)}{x_\phi^{k+1} g_x(x_\phi,z)}\,. \tag{59}$$

These results now allow us to work out $v_i^j(z)$, taking the summation over n in (49) introduces $D_{k,j}^i(z)$, so we get

$$\begin{aligned} v_i^j(z) &= \sum_{l=1}^{\infty}\sum_{k=0}^{\infty}\left(\alpha_j^i\right)^{l-1} t_l^i(k) \sum_{\phi=1}^{M}\frac{-f_j^i(x_\phi,z)}{x_\phi^{k+1}g_x(x_\phi,z)} \\ &= \frac{1}{1-\alpha_j^i}\sum_{\phi=1}^{M}\frac{-f_j^i(x_\phi,z)}{x_\phi g_x(x_\phi,z)} T^{i,j}\left(\frac{1}{x_\phi}\right), \end{aligned} \tag{60}$$

where $T^{i,j}(z)$ is defined as

$$T^{i,j}(z) \triangleq \sum_{l=1}^{\infty}(1-\alpha_j^i)\left(\alpha_j^i\right)^{l-1} T_l^i(z)\,. \tag{61}$$

Using (15), (16) and (43), we can rewrite the expression for $T^{i,j}(z)$ as

$$\begin{aligned} T^{i,j}(z) &= \sum_{l=1}^{\infty}(1-\alpha_j^i)\left(\alpha_j^i\right)^{l-1}F_i(z)S_i\left(\frac{1}{z}\right)[Y_i(z)]^{l-1}G_0^i(z) \\ &\quad + \sum_{l=1}^{\infty}(1-\alpha_j^i)\left(\alpha_j^i\right)^{l-1}F_i(z)\sum_{m=0}^{l-1}[Y_i(z)]^m\sum_{k=0}^{\infty}\sum_{n=1}^{\infty}\text{Prob}\left[g_{l-1-m}^i=k\right]s_i(k+n)(1-z^{-n}) \\ &= (1-\alpha_j^i)\frac{F_i(z)S_i\left(\frac{1}{z}\right)}{1-\alpha_j^iY_i(z)}G_0^i(z) + F_i(z)\left[\widehat{Q}_{i,j}(1)-\widehat{Q}_{i,j}\left(\frac{1}{z}\right)\right], \end{aligned} \tag{62}$$

where the functions $\widehat{Q}_{i,j}(z)$ are still unknown, but they are similar to the functions $Q_i(x,z)$ in (18). It can be proven that they are rational functions with denominator equal to the denominator of $S_i(z)$ and that the unknown coefficients of their numerators can be determined using the properties of pgfs. However, it will turn out that it is not necessary to determine these unknowns.

We now have an expression for $W(z)$, the pgf of the delay of an arbitrary customer:

$$W(z) = \sum_i \frac{h^i}{\overline{r}_i}\left\{\sum_{j=1}^{M_i}\eta_j^i\sum_{l=1}^{j}u_i^{j,l}(z) + \sum_{j=1}^{N_i}\omega_j^i v_i^j(z)\right\}. \tag{63}$$

The above expression is rather complex and, therefore, cannot be easily inverted to give the full delay analysis. On top of that, it still contains (a finite number of) unknowns. However, a tail approximation can be obtained, which will be done in the following section.

6. Tail Approximation

To obtain a tail approximation of the delay, we can use the theory of the dominant singularity, which has been used extensively in the literature, see for example [11,25]. The theory stipulates that

$$\text{Prob[Delay } = k \text{ slots]} \approx \frac{-w_0}{z_0}z_0^{-k}; \tag{64}$$

$$\text{Prob[Delay } > k \text{ slots]} \approx \frac{-w_0}{z_0(z_0-1)}z_0^{-k}, \tag{65}$$

for k sufficiently large and with z_0 the pole of $W(z)$ with smallest modulus and with

$$w_0 \triangleq \lim_{z\to z_0} W(z)(z-z_0). \tag{66}$$

Note that z_0 will be positive, real-valued and strictly larger than 1, see, e.g., [11]. Let us take a closer look at $W(z)$ and its subparts (52) and (60) to determine where the dominant pole z_0 can be found. The functions $f^i(x,z)$ are polynomials and, therefore, contain no poles. The x_ϕ ($\phi = 1,\ldots,M$) are assumed to be single roots of $g(x,z)$ and, therefore, $g_x(x_\phi,z)$ contains no zeros. Furthermore, $x_\phi = 0$ cannot give a pole, as $g(0,z) = 0$ has no solutions. Therefore, z_0 can only be a pole of $S_i(x_\phi)$, a pole of $T_l^i\left(\frac{1}{x_\phi}\right)$ or a pole of $T^{i,j}\left(\frac{1}{x_\phi}\right)$.

Conjecture 1. *The dominant pole z_0 can only be found as a pole of $G_0^i\left(\frac{1}{x_\phi}\right)$.*

From (17) it follows that $G_0^1\left(\frac{1}{x_\phi}\right)$ and $G_0^2\left(\frac{1}{x_\phi}\right)$ have the same poles. The dominant pole z_0, therefore, appears in $T_l^i\left(\frac{1}{x_\phi}\right)$ and $T^{i,j}\left(\frac{1}{x_\phi}\right)$, for $i = 1,2$ and $j = 1,\ldots,N_i$. It is found for a specific value of ϕ, which we will call ξ and we will denote the value of $x_\xi(z_0)$ as x_0. Using (66), we can find an expression for w_0:

$$w_0 = \sum_i \frac{h^i}{\bar{r}_i} \left\{ \sum_{j=1}^{M_i} \eta_j^i \sum_{l=1}^{j} u_i^{j,l*}(z_0) + \sum_{j=1}^{N_i} \omega_j^i v_i^{j*}(z_0) \right\}, \tag{67}$$

with

$$u_i^{j,l*}(z_0) \triangleq S_{\bar{\imath}}(x_0)^{j-l} S_{\bar{\imath}}(x_0) \frac{-z_0^{j-l+1} f^{\bar{\imath}}(x_0, z_0)}{x_0 g_x(x_0, z_0)} T_l^{i*}\left(\frac{1}{x_0}\right); \tag{68}$$

$$v_i^{j*}(z_0) \triangleq \frac{1}{1-\alpha_j^i} \frac{-f_j^i(x_0, z_0)}{x_0 g_x(x_0, z_0)} T^{i,j*}\left(\frac{1}{x_0}\right), \tag{69}$$

where furthermore

$$T_l^{i*}\left(\frac{1}{x_0}\right) \triangleq F_i\left(\frac{1}{x_0}\right) S_{\bar{\imath}}(x_0) Y_i\left(\frac{1}{x_0}\right)^{l-1} G_0^{i*}\left(\frac{1}{x_0}\right); \tag{70}$$

$$T^{i,j*}\left(\frac{1}{x_0}\right) \triangleq (1-\alpha_j^i) \frac{F_i\left(\frac{1}{x_0}\right) S_{\bar{\imath}}(x_0)}{1-\alpha_j^i Y_i\left(\frac{1}{x_0}\right)} G_0^{i*}\left(\frac{1}{x_0}\right), \tag{71}$$

and

$$G_0^{i*}\left(\frac{1}{x_0}\right) = \lim_{z \to z_0} G_0^i\left(\frac{1}{x_\xi(z)}\right)(z - z_0). \tag{72}$$

Following an application of L'Hôpital's Rule $G_0^{i*}\left(\frac{1}{x_0}\right)$ is then obtained from (17) by dividing the numerator of $G_0^i\left(\frac{1}{x_\xi(z)}\right)$ by the derivative with respect to z of its denominator and evaluating at $z = z_0$:

$$G_0^{i*}\left(\frac{1}{x_0}\right) \triangleq x_0^2 \frac{S_{\bar{\imath}}(x_0) R_{\bar{\imath}}\left(Y_{\bar{\imath}}\left(\frac{1}{x_0}\right)\right)\left[Q_i\left(Y_i\left(\frac{1}{x_0}\right), 1\right) - Q_i\left(Y_i\left(\frac{1}{x_0}\right), x_0\right)\right] + S_i(x_0)\left[Q_{\bar{\imath}}\left(Y_{\bar{\imath}}\left(\frac{1}{x_0}\right), 1\right) - Q_{\bar{\imath}}\left(Y_{\bar{\imath}}\left(\frac{1}{x_0}\right), x_0\right)\right]}{S_i(x_0) S_{\bar{\imath}}(x_0)\left[R_i'\left(Y_i\left(\frac{1}{x_0}\right)\right) Y_i'\left(\frac{1}{x_0}\right) R_{\bar{\imath}}\left(Y_{\bar{\imath}}\left(\frac{1}{x_0}\right)\right) + R_i\left(Y_i\left(\frac{1}{x_0}\right)\right) R_{\bar{\imath}}'\left(Y_{\bar{\imath}}\left(\frac{1}{x_0}\right)\right) Y_{\bar{\imath}}'\left(\frac{1}{x_0}\right)\right] \left.\frac{dx_\xi}{dz}\right|_{z=z_0}}. \tag{73}$$

In order to evaluate $\left.\frac{dx_\xi}{dz}\right|_{z=z_0}$ we recall that x_ξ is a solution of

$$1 - R_1(zS_1(x_\xi)) R_2(zS_2(x_\xi)) = 0. \tag{74}$$

Deriving both sides of the above equation with respect to z, working out for $\frac{dx_\xi}{dz}$ and evaluating at $z = z_0$, we find

$$\left.\frac{dx_\xi}{dz}\right|_{z=z_0} = -\frac{S_1(x_0) R_2(z_0 S_2(x_0)) R_1'(z_0 S_1(x_0)) + S_2(x_0) R_1(z_0 S_1(x_0)) R_2'(z_0 S_2(x_0))}{z_0 S_1'(x_0) R_2(z_0 S_2(x_0)) R_1'(z_0 S_1(x_0)) + z_0 S_2'(x_0) R_1(z_0 S_1(x_0)) R_2'(z_0 S_2(x_0))}. \tag{75}$$

Note that z_0 does not necessarily exist. Indeed, if we consider the case where every slot contains at most 1 arrival, every arriving customer will experience an empty queue, and will be served in the slot following its slot of arrival. The delay will be 1 for all customers, i.e., $W(z) = z$.

7. Numerical Examples

In this section, we will illustrate the model with some numerical examples and validate the obtained results using simulation. First we compare the delay characteristics of two very similar queueing systems and then we look at the influence of the period lengths (while keeping their ratio constant).

7.1. Two Similar Queueing Systems

In this example, we compare two cases with the following input parameters:

Case A	**Case B**
$C_1(z) = e^{\lambda_1(z-1)}$; $\lambda_1 = 0.1$	$C_1(z) = e^{\lambda_1(z-1)}$; $\lambda_1 = 4.25$
$C_2(z) = e^{\lambda_2(z-1)}$; $\lambda_2 = 1.985$	$C_2(z) = e^{\lambda_2(z-1)}$; $\lambda_2 = 0.1$
$S_1(z) = \dfrac{20z - 19z^2}{44 - 64z + 21z^2}$; $\bar{s}_1 = 4$	$S_1(z) = \dfrac{20z - 19z^2}{44 - 64z + 21z^2}$; $\bar{s}_1 = 4$
$S_2(z) = \dfrac{1.5z - 0.5z^2}{2 - z}$; $\bar{s}_2 = 1.5$	$S_2(z) = \dfrac{1.5z - 0.5z^2}{2 - z}$; $\bar{s}_2 = 1.5$
$R_1(z) = \dfrac{1}{10 - 9z}$; $\bar{r}_1 = 10$	$R_1(z) = \dfrac{1}{10 - 9z}$; $\bar{r}_1 = 10$
$R_2(z) = \dfrac{1}{20 - 19z}$; $\bar{r}_2 = 20$	$R_2(z) = \dfrac{1}{20 - 19z}$; $\bar{r}_2 = 20$

For the stochastic processes describing the numbers of servers available per slot, and the period lengths, we take the same distributions for both Case A and Case B. The expected numbers of available servers are $\bar{s}_1 = 4$ and $\bar{s}_2 = 1.5$, and the expected period lengths are $\bar{r}_1 = 10$ and $\bar{r}_2 = 20$. For the numbers of arrivals per slot we take Poisson distributions with intensities as given above. The chosen values mean that in Case A in the first state there are barely any arrivals and in the second state the system is overloaded, while in Case B this situation is reversed.

This specific choice for the parameters results in two queueing systems with an equal expected delay of 10.0 slots, which can be obtained using the method of [16] and using Little's Law [26], and which was also confirmed by simulation. Using the method of the current paper, it can be seen that despite having the same expected delay, and despite being two very similar queueing systems, the tail characteristics of the delay are different, as plotted in Figure 2. We observe that for Case B, there is a larger probability that the delay exceeds a large given value, i.e., the delay distribution of customers has a heavier tail in Case B as compared to Case A.

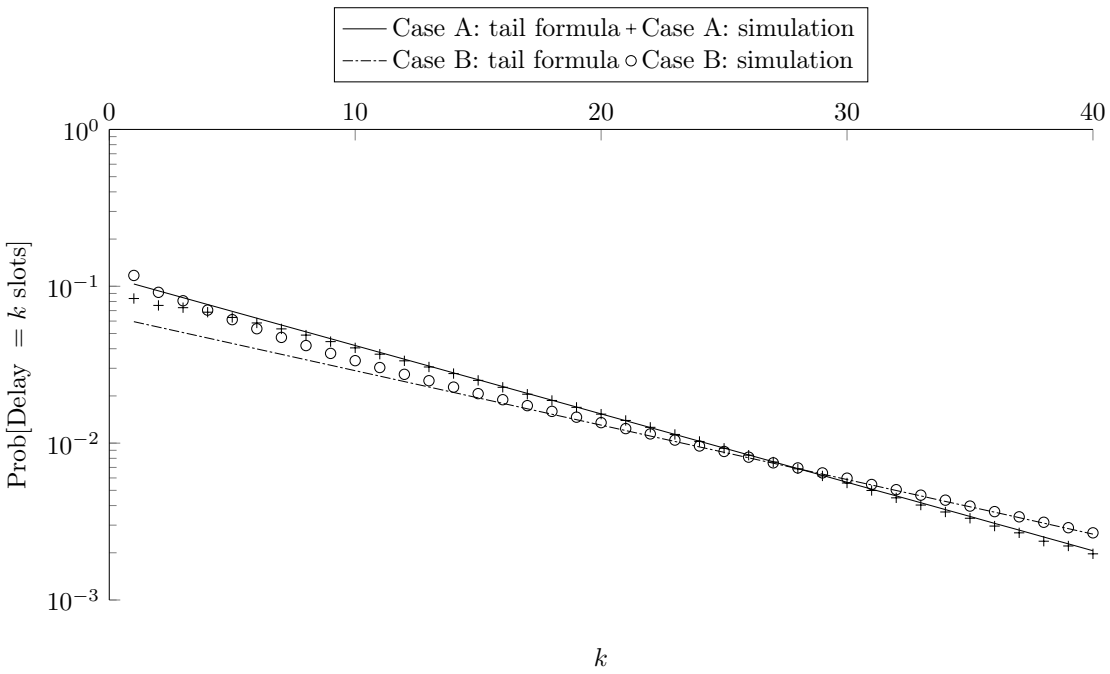

Figure 2. Distribution of delay for Case A and Case B both based on the theory of this paper (tail distribution) and based on simulation (probability mass function).

In the figure the probability mass function of the delay obtained by simulation is also plotted, as a validation of the tail probabilities obtained via the method developed in this paper. For Case A, the results obtained by the method of this paper are within 5% of the results obtained by simulation for $k > 8$. For Case B, this is so for $k > 17$. The difference between the simulated and calculated results decreases for increasing k, e.g., for Case A the difference is under 1% for $k > 15$. The method of this paper thus clearly provides the tail probabilities of the delay with a good accuracy for k sufficiently large. This was verified in several other settings as well, with similar observations on the accuracy. It must be noted that the computational effort to obtain the results by simulation is much higher compared to the method of this paper.

7.2. Influence of Period Lengths

In a second example, we look at the influence of the period lengths on the average delay and 99th percentile. Throughout this subsection, we take the arrival process and the distribution of the period lengths the same for both states, namely a Poisson arrival process with intensity λ and a geometrical distribution for the period lengths with parameter $1 - \frac{1}{\bar{r}}$ (and thus an expected period length of $\bar{r}$):

$$C_1(z) = C_2(z) = e^{\lambda(z-1)}; \tag{76}$$

$$R_1(z) = R_2(z) = \frac{z}{\bar{r} - (\bar{r}-1)z}. \tag{77}$$

For the server availability, we take a fixed number of four servers available per slot in the first state and a fixed number of two servers available per slot in the second state:

$$S_1(z) = z^4; \quad S_2(z) = z^2. \tag{78}$$

In Figure 3, we plot the average delay and the 99th percentile of the delay in function of the expected period lengths $\bar{r}$ for $\lambda = 2.4$. The average delay is calculated from the average system content (as obtained by the method of [16]) by application of Little's Law [26]. The 99th percentile of the delay, i.e., the smallest k for which Prob[delay $> k$ slots] < 0.01, is calculated by inversion of (65):

$$\text{99th percentile } = \text{ceil}\left(-\frac{\ln\left(\frac{0.01 \cdot z_0(z_0-1)}{-w_0}\right)}{\ln(z_0)} \right). \tag{79}$$

As, on average, there are three servers available, the load is 80%, but in the second state, the system is temporarily overloaded. With increasing period lengths, the average delay increases linearly. This is to be expected as the system remains for longer periods in the overloaded state. The same trend is visible for the 99th percentile of the delay; however, the increase happens with a much steeper slope. This proves that valuable information can be obtained from a tail analysis of the delay. The results of Figure 3 were verified by simulation as well (not plotted as they are not distinguishable from the results obtained by the method of this paper).

7.3. Brief Summary for Implementation

In this subsection, we briefly indicate how the method of this paper can be implemented in order to obtain results such as for the numerical examples given earlier in this section, starting from the pgfs fully describing the queueing system ($R_i(z)$, $S_i(z)$ and $C_i(z)$, as defined in Section 2).

1. Numerically obtain the solutions z_p of

$$1 - R_1(Y_1(z))R_2(Y_2(z)) = 0 \quad \text{for } |z| \leq 1, \tag{80}$$

where $Y_i(z) = C_i(z)S_i\left(\frac{1}{z}\right)$ and with $z_1 = 1$. There are a total of $M = m_r^1 m_s^1 + m_r^2 m_s^2$ such solutions.

2. Introduce the functions $Q_i(x,z)$ according to (18). Calculate the unknowns ϵ_{nj}^i by solving the following set of equations for $i = 1$ or for $i = 2$ (both give an equivalent set of equations):

$$S_{\bar{i}}\left(\frac{1}{z_p}\right) R_{\bar{i}}(Y_{\bar{i}}(z_p)) \left[Q_i(Y_i(z_p),1) - Q_i\left(Y_i(z_p), \frac{1}{z_p}\right)\right] + S_i\left(\frac{1}{z_p}\right) \left[Q_{\bar{i}}(Y_{\bar{i}}(z_p),1) - Q_{\bar{i}}\left(Y_{\bar{i}}(z_p), \frac{1}{z_p}\right)\right] = 0\,, \quad p = 2 \ldots M\,; \tag{81}$$

$$\frac{\frac{\partial}{\partial z} Q_1(x,z)|_{(x,z)=(1,1)} + \frac{\partial}{\partial z} Q_2(x,z)|_{(x,z)=(1,1)}}{\bar{r}_1(\bar{s}_1 - \lambda_1) + \bar{r}_2(\bar{s}_2 - \lambda_2)} = 1\,. \tag{82}$$

3. Numerically obtain the solutions x_p of

$$1 - R_1\left(Y_1\left(\frac{1}{x}\right)\right) R_2\left(Y_2\left(\frac{1}{x}\right)\right) = 0\,; \quad |x| < 1. \tag{83}$$

4. Find z_0 as the solution closest to 1 of the following polynomial Equation (for all possible x_p):

$$g(x_p, z) = 0\,, \tag{84}$$

with $g(x,z)$ given in (36) and x_0 is then equal to the corresponding x_p.

5. Fill in the obtained values for z_0 and x_0 in the expression (67) for w_0.

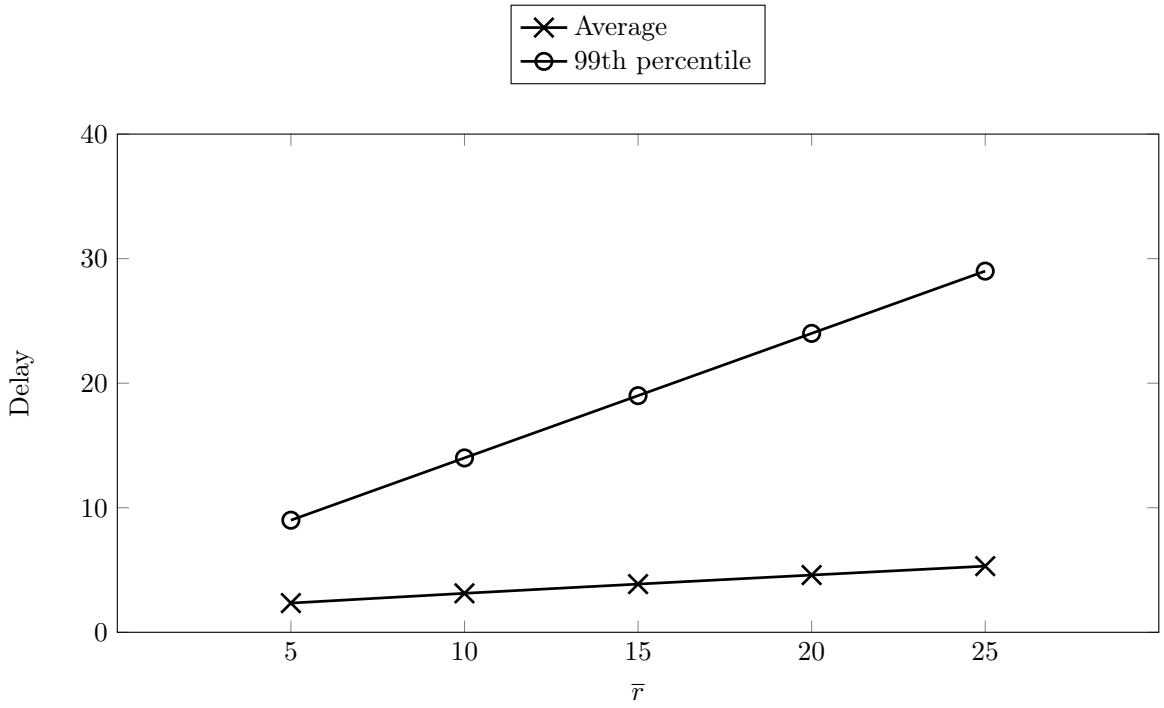

Figure 3. Average and 99th percentile of the delay for the example of Section 7.2.

8. Conclusions

In this paper, we have studied the delay characteristics of a discrete-time multiserver queueing model. It is the extension of an earlier work where the system content was treated [16].

The queueing model considers two different system states. Each state is characterized by its own distributions for the number of arrivals and the number of available servers in a slot. Within a state, these numbers are independent and identically distributed random variables. The state transitions occur after stochastic period lengths, and each state has its own distribution for the sojourn time in that state. This setup allows to model a broad variety of queueing systems, where correlations can be introduced in the slot-to-slot arrivals and server availability.

In this paper, we have obtained the tail distribution of the delay for such a queueing model, using a generating functions approach and using the theory of the dominant singularity. Numerical examples have shown the accuracy of the obtained results, and the importance of the work. We have seen that tail characteristics of the delay of a customer in two separate queueing systems can be substantially different even when the systems are very similar and the expected delay is the same. Tail characteristics also provide a further in-depth understanding of the behavior of a queueing system.

Author Contributions: Conceptualization, F.V., H.B. and S.W.; methodology, F.V., H.B. and S.W.; validation, F.V. and S.W.; formal analysis, F.V.; investigation, F.V.; data curation, F.V.; writing—original draft preparation, F.V.; writing—review and editing, F.V. and S.W.; visualization, F.V.; supervision, H.B. and S.W.; funding acquisition, H.B. and S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Bruneel, H.; Kim, B.G. *Discrete-Time Models for Communication Systems Including ATM*; Kluwer Academic Publishers Group: Amsterdam, The Netherlands, 1993.
2. Daigle, J. *Queueing Theory with Applications to Packet Telecommunication*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005.
3. Alfa, A. *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
4. Kiefer, J.; Wolfowitz, J. On the theory of queues with many servers. *Trans. Am. Math. Soc.* **1955**, *78*, 1–18. [CrossRef]
5. Kim, J.; Kim, B.; Kang, J. Discrete-time multiserver queue with impatient customers. *Electron. Lett.* **2013**, *49*, 38–39.
6. Chaudhry, M.L.; Kim, J.J.; Banik, A.D. Analytically simple and computationally efficient results for the GI(X)/Geo/c queues. *J. Probab. Stat.* **2019**, *2019*, 6480139. [CrossRef]
7. Georganas, N.D. Buffer behavior with Poisson arrivals and bulk geometric service. *IEEE Trans. Commun.* **1976**, *24*, 938–940. [CrossRef]
8. Bruneel, H. A discrete-time queueing system with a stochastic number of servers subjected to random interruptions. *Opsearch* **1985**, *22*, 215–231.
9. Goswami, V.; Mund, G. Computational analysis of multi-server discrete-time queueing system with balking, reneging and synchronous vacations. *RAIRO Oper. Res.* **2017**, *51*, 343–358. [CrossRef]
10. Chaudhry, M.L.; Kim, N.K. A complete and simple solution for a discrete-time multi-server queue with bulk arrivals and deterministic service times. *Oper. Res. Lett.* **2003**, *31*, 101–107. [CrossRef]
11. Bruneel, H.; Steyaert, B.; Desmet, E.; Petit, G. Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues. *Eur. J. Oper. Res.* **1994**, *76*, 563–572. [CrossRef]
12. Gao, P.; Wittevrongel, S.; Bruneel, H. Discrete-time multiserver queues with geometric service times. *Comput. Oper. Res.* **2004**, *31*, 81–99. [CrossRef]
13. Laevens, K.; Bruneel, H. Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. *Eur. J. Oper. Res.* **1995**, *85*, 161–177. [CrossRef]
14. Neuts, M.; Lucantoni, D. A Markovian queue with N servers subject to breakdowns and repair. *Manag. Sci.* **1979**, *25*, 849–861. [CrossRef]
15. De Muynck, M.; Bruneel, H.; Wittevrongel, S. Analysis of a queue with general service demands and correlated service capacities. *Ann. Oper. Res.* **2020**, *293*, 73–99. [CrossRef]

16. Verdonck, F.; Bruneel, H.; Wittevrongel, S. Analysis of a 2-state discrete-time queue with stochastic state-period lengths and state-dependent server availability and arrivals. *Perform. Eval.* **2019**, *135*. [CrossRef]
17. Ibrahim, R.; Whitt, W. Wait-time predictors for customer service systems with time-varying demand and capacity. *Oper. Res.* **2011**, *59*, 1106–1118. [CrossRef]
18. Liu, Y.; Whitt, W. The G t/GI/s t+GI many-server fluid queue. *Queueing Syst.* **2012**, *71*, 405–444. [CrossRef]
19. Grier, N.; Massy, W.; McKoy, T.; Whitt, W. The time-dependent Erlang loss model with retrials. *Telecommun. Syst.* **1997**, *7*, 253–265. [CrossRef]
20. Massey, W. The analysis of queues with time-varying rates for telecommunication models. *Telecommun. Syst.* **2002**, *21*, 173–204. [CrossRef]
21. Wierman, A.; Andrew, L.L.; Tang, A. Power-aware speed scaling in processor sharing systems. In Proceedings of the IEEE INFOCOM, Rio de Janeiro, Brazil, 19–25 April 2009.
22. Tipper, D.; Yi, Q.; Hou, X. Modeling the time varying behavior of mobile ad hoc networks. In Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Venice, Italy, 4–6 October 2004; pp. 12–19.
23. Kleinrock, L. *Queueing Systems, Volume I: Theory*; Wiley: Hoboken, NJ, USA, 1975.
24. Takagi, H. *Queueing Analysis—A Foundation of Performance Evaluation, Volume 3: Discrete-Time Systems*; North Holland Amsterdam: Amsterdam, The Netherlands, 1993.
25. Woodside, C.; Ho, E. Engineering calculation of overflow probabilities in buffers with Markov-interrupted service. *IEEE Trans. Commun.* **1987**, *35*, 1272–1277. [CrossRef]
26. Little, J. A proof for the queueing formula: L = λ W. *Oper. Res.* **1961**, *9*, 383–387. [CrossRef]