


An All Geometric Discrete-Time Multiserver Queueing System

Freek Verdonck ^[0000-0002-8889-1009], Herwig Bruneel^[0000-0002-3739-327X],
and Sabine Wittevrongel^[0000-0001-6985-8361]

Ghent University (UGent), Department of Telecommunications and Information Processing (TELIN), SMACS Research Group, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

`{freek.verdonck,herwig.bruneel,sabine.wittevrongel}@ugent.be`

Abstract. In this work we look at a discrete-time multiserver queueing system where the number of available servers is distributed according to one of two geometrics. The arrival process is assumed to be general independent, the service times deterministically equal to one slot and the buffer capacity infinite. The queueing system resides in one of two states and the number of available servers follows a geometric distribution with parameter determined by the system state. At the end of a slot there is a fixed probability that the system evolves from one state to the other, with this probability depending on the current system state only, resulting in geometrically distributed sojourn times.

We obtain the probability generating function (pgf) of the system content of an arbitrary slot in steady-state, as well as the pgf of the system content at the beginning of an arbitrary slot with a given state. Furthermore we obtain an approximation of the distribution of the delay a customer experiences in the proposed queueing system. This approximation is validated by simulation and the results are illustrated with a numerical example.

Keywords: Queueing theory · Discrete-time · Multiserver · Geometric · System content · Delay.

1 Introduction

In many queueing situations the number of servers is not constant. Moreover, the expected number of available servers can vary over time. In this respect, this paper focusses on a discrete-time queueing model with two system states. The number of available servers in a slot follows a geometric distribution with a parameter that is determined by the system state. The arrival process is assumed to be general and independent, service times are deterministically equal to one slot and the buffer capacity is infinite. State changes are assumed to occur according to a first-order Markov process.

Discrete-time multiserver queueing systems have received considerable attention in the past in several settings [5,6,7,8,13]. Specifically, [13] handles a

multiserver queueing system with priorities. Multiserver queueing systems with batch arrivals are considered in [5,6] for geometric, respectively deterministic service times, while [7,8] deal with multiserver systems with general independent arrivals and geometric and constant service times. These papers all have in common that the number of servers present in the system is constant.

Also queues with a varying number of servers have been studied in literature. In the most simple case, the number of available servers changes independently from slot to slot [1,9]. In [9] either all m servers are available or none, while in [1] the number of available servers can take any value between 0 and m . In [4] correlation over time is introduced on the number of available servers by combining a permanently available server with an extra server with generally distributed off-times and geometrically distributed on-times. The analysis of [4], however, is limited to the system content. In the current paper, we focus on a discrete-time multiserver queueing model that is general enough to include variability and time correlation on the number of available servers, but yet simple enough to lend itself to a full queueing analysis of not only the system content but also the delay, and to have a limited number of easily interpretable model parameters. Specifically, our model considers two possible system states, each with their own geometric distributions for state sojourn time and for the number of available servers during a slot.

The main contribution of the current paper is the delay analysis for the considered queueing model. Multiserver queueing systems are notoriously hard when considering the delay analysis, especially when dealing with a varying number of available servers. Some earlier results can be found in [12] where all m servers are subject to independent interruptions and no correlation is present on the server availability.

The study of this model is motivated by the many applications of queueing theory where the number of servers is not constant over time and a certain correlation exists in the number of available servers. Examples include the modelling of the airport checkin process [14] or supply chains in production facilities [11].

The outline of the paper is as follows. In the next section we provide a detailed mathematical description of the queueing model under study. In Section 3 we obtain the steady-state distribution of the system content at the beginning of an arbitrary slot and at the beginning of a slot with a given state. In Section 4 we look at the delay analysis for this queueing system. Section 5 provides a numerical example and Section 6 concludes the paper.

2 Mathematical Model

In this paper we study a discrete-time queueing system; the time horizon is divided into slots of equal length. The arrival process is general and independent and is described by

$$c(n) \triangleq \text{Prob}[n \text{ customers arrive during a slot}] , \quad n \geq 0; \quad (1)$$

$$C(z) \triangleq \sum_{n=0}^{\infty} c(n)z^n; \tag{2}$$

$$\lambda \triangleq \sum_{n=0}^{\infty} nc(n) = C'(1) . \tag{3}$$

If we denote by c_k the number of arrivals in the k th slot, then the series $\{c_k\}$ is a set of independent and identically distributed (i.i.d.) stochastic variables. The service time of every customer equals 1 slot.

The system resides in state-A or state-B and the number of available servers during a slot follows a geometrical distribution, with parameter determined by the system state during that slot. Specifically, we have that

$$\text{Prob}[n \text{ servers available during A-slot}] = (1 - \beta_1)\beta_1^n, \quad n \geq 0; \tag{4}$$

$$\text{Prob}[n \text{ servers available during B-slot}] = (1 - \beta_2)\beta_2^n, \quad n \geq 0, \tag{5}$$

with $0 < \beta_1, \beta_2 < 1$. The expected number of available servers during an A-slot is $\frac{\beta_1}{1-\beta_1}$ and during a B-slot $\frac{\beta_2}{1-\beta_2}$. We easily obtain

$$\text{Prob}[\text{more than } n \text{ servers available during A-slot}] = \beta_1^{n+1}, \quad n \geq 0; \tag{6}$$

$$\text{Prob}[\text{more than } n \text{ servers available during B-slot}] = \beta_2^{n+1}, \quad n \geq 0. \tag{7}$$

If we denote by $s_{A,k}$ and $s_{B,k}$ the number of available servers in the k th A-slot and the k th B-slot respectively, then we have that the series $\{s_{A,k}\}$ and $\{s_{B,k}\}$ are two different sets of i.i.d. stochastic variables.

State changes can only occur at the end of a slot, and the probability of a state change occurring is fixed and depends solely on the current state:

$$\text{Prob}[\text{A-slot is followed by B-slot}] \triangleq 1 - \alpha_1; \tag{8}$$

$$\text{Prob}[\text{B-slot is followed by A-slot}] \triangleq 1 - \alpha_2. \tag{9}$$

with $0 < \alpha_1, \alpha_2 < 1$. Let us introduce σ to denote the probability that an arbitrary slot is an A-slot, then standard probability theory leads to:

$$\sigma = \frac{1 - \alpha_2}{2 - \alpha_1 - \alpha_2}. \tag{10}$$

This paper handles the steady-state situation of the queueing system as described and it is therefore assumed that the stability condition is fulfilled. For the queue to be stable it is required that the average number of arrivals during a slot is strictly smaller than the average number of customers that can be served. This can be expressed as:

$$\lambda < \frac{\sigma\beta_1}{1-\beta_1} + \frac{(1-\sigma)\beta_2}{1-\beta_2} = \frac{(1-\alpha_1)(1-\beta_1)\beta_2 + (1-\alpha_2)(1-\beta_2)\beta_1}{(1-\beta_1)(1-\beta_2)(2-\alpha_1-\alpha_2)}. \quad (11)$$

We assume a First In First Out (FIFO) policy. The delay of a customer is defined as its total system time, excluding the remainder of the slot in which the customer arrives. The delay is thus always an integer number of slots and includes the service time.

The setup as described is also referred to as a Late Arrival System with Delayed Access (LAS-DA).

A schematic overview of the system can be found in Figure 1.

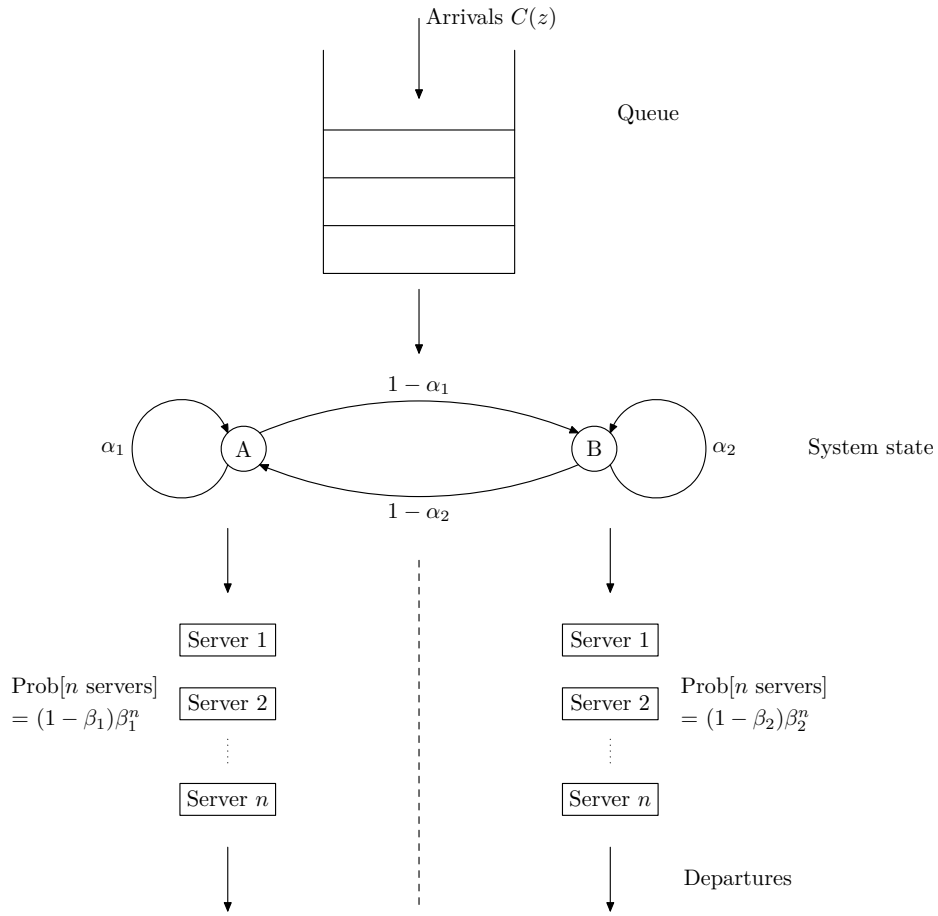


Fig. 1. Illustration of the considered queueing system

3 Analysis of System Content

In this section we first analyze the system content. Let us look at an arbitrary A-slot S_A and an arbitrary B-slot S_B in steady state. We denote the system content at the beginning of these slots as u_A and u_B respectively, while we denote the system content at the end of these arbitrary slots as u_A^+ and u_B^+ respectively. These system contents are related to each other through the following system equations:

$$u_A^+ = \max(u_A - s_A, 0) + c_A; \quad (12)$$

$$u_B^+ = \max(u_B - s_B, 0) + c_B, \quad (13)$$

with s_A and s_B the number of available servers in S_A and S_B , and with c_A and c_B the number of customers arriving in these respective slots. Due to the Markovian transitions between the system states we get that in steady state there is a probability α_1 that the slot before S_A is an A-slot (and a probability $(1 - \alpha_1)$ that the slot before S_A is a B-slot). If we introduce $U_A(z)$ and $U_B(z)$ as the probability generating functions (pgfs) of u_A and u_B respectively, we can write:

$$\begin{aligned} U_A(z) &= \alpha_1 E[z^{u_A^+}] + (1 - \alpha_1) E[z^{u_B^+}] \\ &= \alpha_1 C(z) \sum_{k=0}^{\infty} \text{Prob}[u_A = k] \text{Prob}[s_A > k] \\ &\quad + \alpha_1 C(z) \sum_{k=0}^{\infty} \sum_{l=0}^k \text{Prob}[u_A = k] \text{Prob}[s_A = l] z^{k-l} \\ &\quad + (1 - \alpha_1) C(z) \sum_{k=0}^{\infty} \text{Prob}[u_B = k] \text{Prob}[s_B > k] \\ &\quad + (1 - \alpha_1) C(z) \sum_{k=0}^{\infty} \sum_{l=0}^k \text{Prob}[u_B = k] \text{Prob}[s_B = l] z^{k-l} \\ &= \alpha_1 C(z) \beta_1 U_A(\beta_1) + (1 - \alpha_1) C(z) \beta_2 U_B(\beta_2) \\ &\quad + \alpha_1 C(z) \sum_{k=0}^{\infty} \text{Prob}[u_A = k] z^k (1 - \beta_1) \frac{1 - \left(\frac{\beta_1}{z}\right)^{k+1}}{1 - \frac{\beta_1}{z}} \\ &\quad + (1 - \alpha_1) C(z) \sum_{k=0}^{\infty} \text{Prob}[u_B = k] z^k (1 - \beta_2) \frac{1 - \left(\frac{\beta_2}{z}\right)^{k+1}}{1 - \frac{\beta_2}{z}} \\ &= \alpha_1 C(z) U_A(\beta_1) \beta_1 \frac{z-1}{z-\beta_1} + (1 - \alpha_1) C(z) U_B(\beta_2) \beta_2 \frac{z-1}{z-\beta_2} \\ &\quad + \alpha_1 C(z) U_A(z) \frac{(1-\beta_1)z}{z-\beta_1} + (1 - \alpha_1) C(z) U_B(z) \frac{(1-\beta_2)z}{z-\beta_2}. \quad (14) \end{aligned}$$

A similar derivation can be made for the pgf $U_B(z)$ of u_B leading to

$$U_B(z) = \alpha_2 C(z) U_B(\beta_2) \beta_2 \frac{z-1}{z-\beta_2} + (1-\alpha_2) C(z) U_A(\beta_1) \beta_1 \frac{z-1}{z-\beta_1} \\ + \alpha_2 C(z) U_B(z) \frac{(1-\beta_2)z}{z-\beta_2} + (1-\alpha_2) C(z) U_A(z) \frac{(1-\beta_1)z}{z-\beta_1}. \quad (15)$$

The set of linear equations (14) and (15) can be solved for $U_A(z)$ and $U_B(z)$ which leads to the following explicit expressions:

$$U_A(z) = \frac{(z-1)C(z) \left\{ \begin{array}{l} \beta_2(1-\alpha_1)(z-\beta_1)U_B(\beta_2) + \beta_1\alpha_1(z-\beta_2)U_A(\beta_1) \\ + \beta_1(1-\beta_2)(1-\alpha_1-\alpha_2)zC(z)U_A(\beta_1) \end{array} \right\}}{(z-\beta_1)(z-\beta_2) - (1-\beta_1)(1-\beta_2)(1-\alpha_1-\alpha_2)z^2C(z)^2 \\ - [\alpha_1(1-\beta_1)(z-\beta_2) + \alpha_2(1-\beta_2)(z-\beta_1)]zC(z)}; \quad (16)$$

$$U_B(z) = \frac{(z-1)C(z) \left\{ \begin{array}{l} \beta_1(1-\alpha_2)(z-\beta_2)U_A(\beta_1) + \beta_2\alpha_2(z-\beta_1)U_B(\beta_2) \\ + \beta_2(1-\beta_1)(1-\alpha_1-\alpha_2)zC(z)U_B(\beta_2) \end{array} \right\}}{(z-\beta_1)(z-\beta_2) - (1-\beta_1)(1-\beta_2)(1-\alpha_1-\alpha_2)z^2C(z)^2 \\ - [\alpha_1(1-\beta_1)(z-\beta_2) + \alpha_2(1-\beta_2)(z-\beta_1)]zC(z)}; \quad (17)$$

in which two unknown constants appear, $U_A(\beta_1)$ and $U_B(\beta_2)$. These unknowns cannot be straightforwardly determined since the substitutions $z = \beta_1$ in (16) and $z = \beta_2$ in (17) lead to two identities. However, we can determine them by relying on the properties of pgfs, namely that they are analytical within the complex unit disk and normalized.

Let us first take a look at the denominator of $U_A(z)$ and $U_B(z)$. It can easily be seen that the first part, $(z-\beta_1)(z-\beta_2)$, has exactly 2 zeros within the complex unit disk. By application of Rouché's theorem we can conclude that the whole denominator also has 2 zeros within the complex unit disk (for more information on Rouché's theorem, see for example [10]). It can easily be verified that $z = 1$ is one of these zeros. Let us call the other zero z_1 . As $U_A(z)$ and $U_B(z)$ are pgfs, they cannot have singularities within the complex unit disk and thus their numerators must also vanish at $z = 1$ and $z = z_1$. This is obviously the case for $z = 1$, irrespectively of the unknown constants. Expressing that the numerator must also vanish for $z = z_1$ leads to one relation linking the 2 unknowns. A second relation can be obtained by expressing the normality condition of pgfs:

$$\lim_{z \rightarrow 1} U_A(z) = 1. \quad (18)$$

After applying L'Hôpital's rule and using (10), we obtain

$$\frac{\sigma\beta_1(1-\beta_2)U_A(\beta_1) + (1-\sigma)(1-\beta_1)\beta_2U_B(\beta_2)}{(1-\beta_1)(1-\beta_2)\lambda + \sigma\beta_1(1-\beta_2) + (1-\sigma)(1-\beta_1)\beta_2} = 1. \quad (19)$$

We can thus determine the 2 unknowns in the expressions for $U_A(z)$ and $U_B(z)$ and obtain the pgfs of the system contents at the beginning of an arbitrary A-slot and B-slot. The system content u at the beginning of an arbitrary slot is then determined by its pgf $U(z)$, with

$$U(z) = \sigma U_A(z) + (1 - \sigma)U_B(z) . \quad (20)$$

From the pgf of the system content we can easily obtain the (central) moments of its distribution, leading to the mean and variation of the system content. Also higher order moments can be calculated.

4 Delay Analysis

Now that we have expressions for the pgfs of the system contents at the beginning of an arbitrary A-slot and arbitrary B-slot, we are in a position to study the delay that a customer experiences in the queueing system. First, we condition the delay of a customer on the state of its arrival slot and on the number of customers waiting in the queue in front of it, where we exclude the customers that are receiving service at the moment of arrival. Then, we combine this with the results of the previous section to obtain an expression for the pgf of the delay of an arbitrary customer. Finally, we use the obtained pgf to derive an approximation for the tail probabilities of the delay.

4.1 Delay of a Customer With k Customers Ahead

We look at an arbitrary customer P_A , arriving during an arbitrary A-slot. We introduce the stochastic variable $d_{A,k}$ for its delay, given that there are k customers in front of P_A in the queue at the moment of its arrival (thus excluding the customers receiving service). The corresponding pgf is $D_{A,k}(z)$. Analogously we study the arbitrary customer P_B , arriving during an arbitrary B-slot. Given that there are k customers in the queue in front of P_B , its delay is denoted by the stochastic variable $d_{B,k}$, with corresponding pgf $D_{B,k}(z)$. By looking at the system state and the number of available servers during the slot after the considered customer's arrival slot, we can obtain the following relations (for $k \geq 0$):

$$\begin{aligned} D_{A,k}(z) = & (1 - \alpha_1)z \text{Prob}[s_B > k] + (1 - \alpha_1)z \sum_{l=0}^k D_{B,k-l}(z) \text{Prob}[s_B = l] \\ & + \alpha_1 z \text{Prob}[s_A > k] + \alpha_1 z \sum_{l=0}^k D_{A,k-l}(z) \text{Prob}[s_A = l] ; \quad (21) \\ D_{B,k}(z) = & (1 - \alpha_2)z \text{Prob}[s_A > k] + (1 - \alpha_2)z \sum_{l=0}^k D_{A,k-l}(z) \text{Prob}[s_A = l] \end{aligned}$$

$$+ \alpha_2 z \text{Prob}[s_B > k] + \alpha_2 z \sum_{l=0}^k D_{B,k-l}(z) \text{Prob}[s_B = l], \quad (22)$$

with s_A and s_B the numbers of available servers in an A-slot or B-slot. Let us now introduce some auxiliary functions:

$$D_A(x, z) \triangleq \sum_{k=0}^{\infty} D_{A,k}(z) x^k; \quad (23)$$

$$D_B(x, z) \triangleq \sum_{k=0}^{\infty} D_{B,k}(z) x^k. \quad (24)$$

Using (21) and (22) to work out these definitions we get

$$\begin{aligned} D_A(x, z) &= (1 - \alpha_1) z \sum_{k=0}^{\infty} \beta_2^{k+1} x^k + (1 - \alpha_1) z \sum_{k=0}^{\infty} x^k \sum_{l=0}^k D_{B,k-l}(z) (1 - \beta_2) \beta_2^l \\ &\quad + \alpha_1 z \sum_{k=0}^{\infty} \beta_1^{k+1} x^k + \alpha_1 z \sum_{k=0}^{\infty} x^k \sum_{l=0}^k D_{A,k-l}(z) (1 - \beta_1) \beta_1^l \\ &= \frac{(1 - \alpha_1) \beta_2 z}{1 - \beta_2 x} + \frac{(1 - \alpha_1)(1 - \beta_2) z}{1 - \beta_2 x} D_B(x, z) \\ &\quad + \frac{\alpha_1 \beta_1 z}{1 - \beta_1 x} + \frac{\alpha_1 (1 - \beta_1) z}{1 - \beta_1 x} D_A(x, z), \end{aligned} \quad (25)$$

and in a similar way

$$\begin{aligned} D_B(x, z) &= \frac{(1 - \alpha_2) \beta_1 z}{1 - \beta_1 x} + \frac{(1 - \alpha_2)(1 - \beta_1) z}{1 - \beta_1 x} D_A(x, z) \\ &\quad + \frac{\alpha_2 \beta_2 z}{1 - \beta_2 x} + \frac{\alpha_2 (1 - \beta_2) z}{1 - \beta_2 x} D_B(x, z). \end{aligned} \quad (26)$$

The set of linear equations (25) and (26) can be solved for $D_A(x, z)$ and $D_B(x, z)$. This leads to the following explicit expressions:

$$D_A(x, z) = \frac{f_A(x, z)}{g(x, z)}; \quad (27)$$

$$D_B(x, z) = \frac{f_B(x, z)}{g(x, z)}, \quad (28)$$

with

$$f_A(x, z) \triangleq \beta_1 \beta_2 z x - [(1 - \alpha_1) \beta_2 + (1 - \beta_2)(1 - \alpha_1 - \alpha_2) \beta_1 z - \alpha_1 \beta_1] z; \quad (29)$$

$$f_B(x, z) \triangleq \beta_1 \beta_2 z x - [(1 - \alpha_2) \beta_1 + (1 - \beta_1)(1 - \alpha_1 - \alpha_2) \beta_2 z - \alpha_2 \beta_2] z, \quad (30)$$

and

$$\begin{aligned} g(x, z) \triangleq & -1 + [\alpha_1(1 - \beta_1) + \alpha_2(1 - \beta_2)] z + (1 - \beta_1)(1 - \beta_2)(1 - \alpha_1 - \alpha_2) z^2 \\ & + [(1 - \alpha_1 z) \beta_2 + (1 - \alpha_2 z) \beta_1 + (\alpha_1 + \alpha_2) \beta_1 \beta_2 z] x - \beta_1 \beta_2 x^2. \end{aligned} \quad (31)$$

(32)

We can consider $D_A(x, z)$ and $D_B(x, z)$ as rational functions in x with numerator of degree 1 and denominator of degree 2. A partial fraction expansion can be made based on the poles in x which we denote as x_1 and x_2 and assume to be distinct. We can then rewrite (27) and (28) as

$$D_A(x, z) = \sum_{i=1}^2 \frac{f_A(x_i, z)}{g_x(x_i, z)(x - x_i)}; \quad (33)$$

$$D_B(x, z) = \sum_{i=1}^2 \frac{f_B(x_i, z)}{g_x(x_i, z)(x - x_i)}, \quad (34)$$

with

$$g_x(x, z) \triangleq \frac{\partial}{\partial x} g(x, z). \quad (35)$$

We obtain an expression for $D_{A,k}(z)$ by evaluating the k th derivative with respect to x of $D_A(x, z)$ at $x = 0$:

$$\begin{aligned} D_{A,k}(z) &= \frac{1}{k!} \left. \frac{\partial^k}{\partial x^k} D_A(x, z) \right|_{x=0} \\ &= \sum_{i=1}^2 \frac{-f_A(x_i, z)}{g_x(x_i, z) x_i^{k+1}}. \end{aligned} \quad (36)$$

In a similar way we find the following expression for $D_{B,k}(z)$:

$$D_{B,k}(z) = \sum_{i=1}^2 \frac{-f_B(x_i, z)}{g_x(x_i, z) x_i^{k+1}}. \quad (37)$$

4.2 Delay of an Arbitrary Customer

We consider the arbitrary packet P_A , arriving in the system during the A-slot S_A and we denote the number of customers in the queue at its moment of arrival

by the stochastic variable t_A , with corresponding pgf $T_A(z)$. Upon arrival, the customers waiting in the queue are those that were present in the queueing system at the beginning of S_A , minus those that receive service during S_A and plus those that arrived during S_A , but before the arrival of P_A . The pgf $F(z)$ of this last number of arrivals is well known in the literature, see e.g. [2]:

$$F(z) = \frac{C(z) - 1}{\lambda(z - 1)}. \quad (38)$$

We get for $T_A(z)$:

$$\begin{aligned} T_A(z) &= F(z) \left\{ \sum_{k=0}^{\infty} \text{Prob}[u_A = k] \text{Prob}[s_A > k] \right. \\ &\quad \left. + \sum_{k=0}^{\infty} \text{Prob}[u_A = k] \sum_{l=0}^k \text{Prob}[s_A = l] z^{k-l} \right\} \\ &= F(z) \left\{ \sum_{k=0}^{\infty} \text{Prob}[u_A = k] \beta_1^{k+1} \right. \\ &\quad \left. + \sum_{k=0}^{\infty} \text{Prob}[u_A = k] z^k \sum_{l=0}^k (1 - \beta_1) \left(\frac{\beta_1}{z} \right)^l \right\} \\ &= F(z) \left\{ \beta_1 U_A(\beta_1) + (1 - \beta_1) \sum_{k=0}^{\infty} \text{Prob}[u_A = k] \frac{z^{k+1} - \beta_1^{k+1}}{z - \beta_1} \right\} \\ &= F(z) \frac{(1 - \beta_1)z U_A(z) + (z - 1)\beta_1 U_A(\beta_1)}{z - \beta_1}. \end{aligned} \quad (39)$$

In a similar manner we can define t_B as the queue content seen by an arbitrary customer P_B , arriving during a B-slot. The corresponding pgf $T_B(z)$ is given by

$$T_B(z) = F(z) \frac{(1 - \beta_2)z U_B(z) + (z - 1)\beta_2 U_B(\beta_2)}{z - \beta_2}. \quad (40)$$

The arrival slot of an arbitrary packet P is an A-slot with probability σ and a B-slot with probability $(1 - \sigma)$. We can therefore express the pgf of its delay $W(z)$ as

$$W(z) = \sigma \sum_{k=0}^{\infty} \text{Prob}[t_A = k] D_{A,k}(z) + (1 - \sigma) \sum_{k=0}^{\infty} \text{Prob}[t_B = k] D_{B,k}(z). \quad (41)$$

Substitution of (36) and (37) into the above expression yields

$$W(z) = \sigma \sum_{i=1}^2 \frac{-f_A(x_i, z)}{g_x(x_i, z) x_i} T_A\left(\frac{1}{x_i}\right) + (1 - \sigma) \sum_{i=1}^2 \frac{-f_B(x_i, z)}{g_x(x_i, z) x_i} T_B\left(\frac{1}{x_i}\right). \quad (42)$$

The above equation is fully determined, the x_i are functions of z , but can be easily obtained as they are the roots of a quadratic equation. However, it is not easy to invert this pgf. In the following subsection we use these results to obtain a tail approximation of the delay of an arbitrary customer.

4.3 Tail Approximation

We can use the technique of the dominant singularity, see e.g. [3,15], to obtain the tail distribution of the delay of an arbitrary customer. For sufficiently large k we have that

$$\text{Prob}[\text{delay} = k \text{ slots}] \approx -\frac{w_0}{z_0} z_0^{-k}; \quad (43)$$

$$\text{Prob}[\text{delay} > k \text{ slots}] \approx -\frac{w_0}{z_0(z_0 - 1)} z_0^{-k}, \quad (44)$$

with z_0 the pole of $W(z)$ with the smallest modulus and with

$$w_0 = \lim_{z \rightarrow z_0} [W(z)(z - z_0)]. \quad (45)$$

Note that z_0 is real-valued and larger than 1 and that w_0 is real-valued and negative. As the x_i are non-zero and distinct we have that z_0 can only be found as a pole of $T_A\left(\frac{1}{x_i}\right)$, or equivalently as a pole of $T_B\left(\frac{1}{x_i}\right)$. In view of (16), (38) and (39), z_0 must be found as a pole of $C\left(\frac{1}{x_i}\right)$ or as a zero of

$$\begin{aligned} f(z) = & (1 - \beta_1 x_i)(1 - \beta_2 x_i) - (1 - \beta_1)(1 - \beta_2)(1 - \alpha_1 - \alpha_2) C\left(\frac{1}{x_i}\right)^2 \\ & - [\alpha_1(1 - \beta_1)(1 - \beta_2 x_i) + \alpha_2(1 - \beta_2)(1 - \beta_1 x_i)] C\left(\frac{1}{x_i}\right). \end{aligned} \quad (46)$$

5 Numerical Examples

In this section we illustrate the method developed in this paper with some numerical examples. For the arrivals we take a Poisson process:

$$C(z) = e^{\lambda(z-1)}. \quad (47)$$

In an initial example we take the following values for the parameters: $\alpha_1 = 0.6$, $\alpha_2 = 0.7$, $\beta_1 = 0.4$ and $\beta_2 = 0.6$. The average number of servers available in an arbitrary slot equals 1.14 and thus the system is stable if $\lambda < 1.14$. In Figure 2 we plot the average system content in function of the arrival intensity λ for this system based on the method developed in this paper, as well as based on

simulation. It must be noted that for the simulation the required computation time is much larger than the time required to compute the unknowns $U_A(\beta_1)$ and $U_B(\beta_2)$, which involves finding the root of a non-polynomial function within the complex unit disk and solving a set of 2 linear equations. The system content shows an expected vertical asymptote for $\lambda \rightarrow 1.14$. The simulation validates the results obtained by the method of this paper.

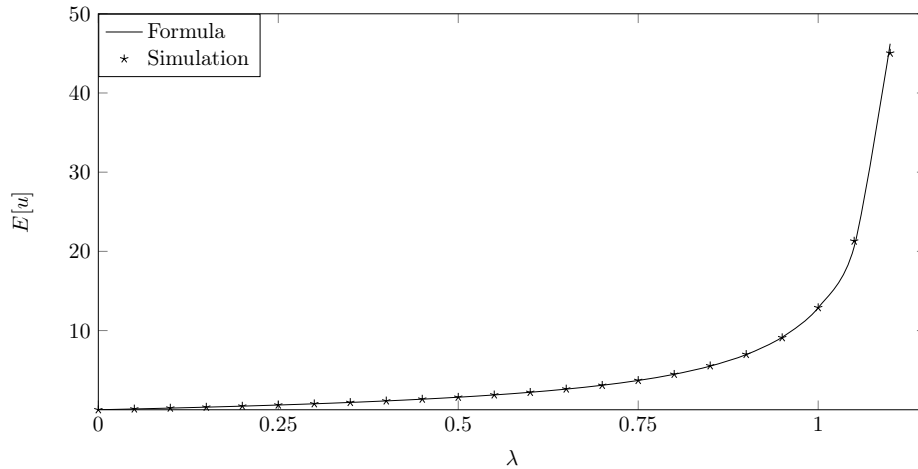


Fig. 2. Average system content in function of λ , for $\alpha_1 = 0.6$, $\alpha_2 = 0.7$, $\beta_1 = 0.4$ and $\beta_2 = 0.6$.

In Figure 3 we look at the delay characteristics for this example. We set $\lambda = 1$ and we plot the probability that the delay of a customer equals k slots for increasing k . The mean system content in this situation equals 12.88 customers. We see that already for small k the obtained tail approximation is very close to the simulation results. For large k the simulation would need to run for a very long time in order to get reliable results, while our formula immediately gives an excellent approximation for all k .

In a second example we introduce a larger difference between the two states: $\alpha_1 = 0.61$, $\alpha_2 = 0.3$, $\beta_1 = 0.1$ and $\beta_2 = 0.75$. The parameters have been chosen in such a way that the expected number of servers available during an arbitrary slot is the same as in the previous example. In the A-state there is now a high probability that no servers are available, while in state-B we expect on average 3 available servers. However, the system does not reside for long periods in the B-state as α_2 is small. In Figure 4 the system content is plotted in function of the arrival intensity λ . Note that the vertical axis now goes until 100. The curve of the system content shows the same shape and has the same vertical asymptote for $\lambda \rightarrow 1.14$, but for a given λ the system content is higher as compared to

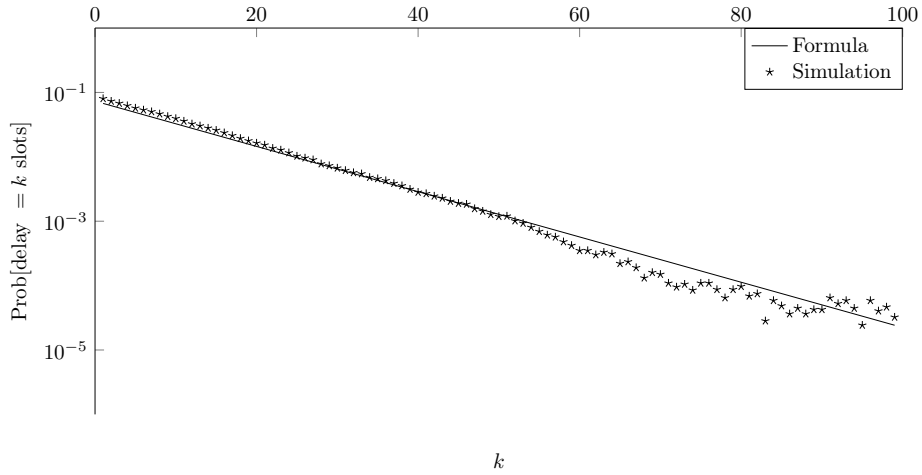


Fig. 3. Delay characteristics for $\lambda = 1$, $\alpha_1 = 0.6$, $\alpha_2 = 0.7$, $\beta_1 = 0.4$ and $\beta_2 = 0.6$.

the first situation. This is expected in view of the increased irregularity in the service process for the second example. The difference is also larger for higher λ .

We now also look at the delay characteristics for this situation. We choose $\lambda = 0.9157$ in order to have the same mean system content of 12.88 as in the previous example. The delay characteristics are plotted in Figure 5. The slope of the delay curve in this situation is less steep than before, the characteristic singularity being closer to 1. In particular we now have $z_0 = 1.0769$, while for Figure 3 we had $z_0 = 1.0843$. This means that even though the mean system content remains the same, the delay characteristics are different. For large values of k , there is a higher probability that the delay of a customer exceeds k as compared to the situation of Figure 3. This is again in accordance with the increased irregularity in the service process.

6 Conclusion

In this paper we have studied an all geometric discrete-time multiserver queueing system. The system resides in one of two different states, and does so for a geometrically distributed number of slots, with a different parameter for each state. The number of servers available during a slot also follows a geometric distribution, with a parameter depending on the system state. The model can be used in many applications of queueing theory where the expected number of available servers fluctuates over time and is fairly simple in terms of the number of parameters that needs to be matched. We have obtained expressions for the probability generating functions of the system content at the beginning of an arbitrary slot and at the beginning of an arbitrary slot with a given state.

Furthermore we have obtained an approximation for the tail probabilities of the delay that an arbitrary customer experiences in the queueing system. This

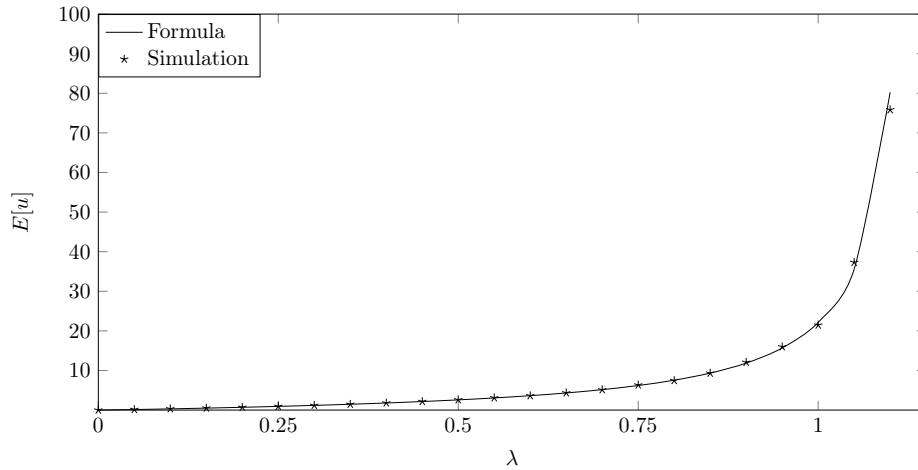


Fig. 4. Average system content in function of λ , for $\alpha_1 = 0.61$, $\alpha_2 = 0.3$, $\beta_1 = 0.1$ and $\beta_2 = 0.75$.

approximation is based on the theory of the dominant singularity. Numerical examples have shown that our tail approximation is also accurate for smaller delay values. The numerical examples have further illustrated that more variation in the number of available servers leads to higher system contents. Moreover, for a given system content, there is a higher probability that a customer experiences larger delays when more variability is present in the system.

References

1. Bruneel, H.: A general model for the behavior of infinite buffers with periodic service opportunities. *European Journal of Operations Research* **16**(1), 98–106 (1984)
2. Bruneel, H., Kim, B.G.: *Discrete-Time Models for Communication Systems Including ATM*. Kluwer Academic Publishers Group (1993)
3. Bruneel, H., Steyaert, B., Desmet, E., Petit, G.: Analytic derivation of tail probabilities for queue lengths and waiting times in atm multiserver queues. *European Journal of Operational Research* **76**(3), 563–572 (August 1994)
4. Bruneel, H., Wittevrongel, S.: Analysis of a discrete-time single-server queue with an occasional extra server. *Performance Evaluation* **116**, 119–142 (2017)
5. Chaudhry, M.L., Gupta, U., Goswami, V.: Modeling and Analysis of Discrete-Time Multiserver Queues with Batch Arrivals: $GI^X/Geom/m$. *INFORMS Journal on Computing*
6. Chaudhry, M.L., Kim, N.: A complete and simple solution for a discrete-time multi-server queue with bulk arrivals and deterministic service times. *Operations Research Letters* **31**(2), 101–107 (2003)
7. Gao, P., Wittevrongel, S., Bruneel, H.: Discrete-time multiserver queues with geometric service times. *Computers & Operations Research* **31**(1), 81–99 (2004)

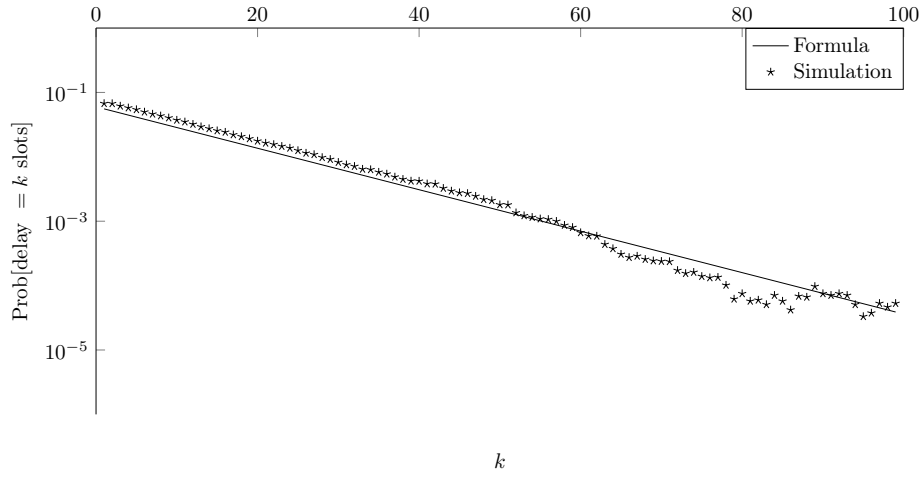


Fig. 5. Delay characteristics for $\lambda = 0.9157$, $\alpha_1 = 0.61$, $\alpha_2 = 0.3$, $\beta_1 = 0.1$ and $\beta_2 = 0.75$.

8. Gao, P., Wittevrongel, S., Walraevens, J., Moeneclaey, M., Bruneel, H.: Calculation of delay characteristics for multiserver queues with constant service times. *European Journal of Operational Research* **199**(1), 170–175 (2009)
9. Georganas, N.D.: Buffer behavior with poisson arrivals and bulk geometric service. *IEEE Transactions on Communications* **24**, 938–940 (1976)
10. Gonzalez, M.: *Classical Complex Analysis*. CRC Press (1991)
11. Kerbache, L., Smith, J.: Queueing networks and the topological design of supply chain systems **91**, 251–272 (2004)
12. Laevens, K., Bruneel, H.: Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. *European Journal of Operations Research* **85**, 161–177 (1995)
13. Laevens, K., Bruneel, H.: Discrete-time multiserver queues with priorities. *Performance evaluation* **33**, 249–275 (1998)
14. Stolletz, R.: *Analysis of passenger queues at airport terminals*. Research in Transportation Business & Management (2011)
15. Woodside, C., Ho, E.: Engineering calculation of overflow probabilities in buffers with markov-interrupted service. *IEEE Transactions on Communications* **35**(12), 1272–1277 (December 1987)