JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **13**, Number **4**, October **2017**

doi:10.3934/jimo.2017024

pp. **1901–1926**

ANALYSIS OF A DISCRETE-TIME QUEUE WITH GENERAL SERVICE DEMANDS AND PHASE-TYPE SERVICE CAPACITIES

Michiel De Muynck*, Herwig Bruneel and Sabine Wittevrongel

SMACS Research Group
Department of Telecommunications and Information Processing
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

Abstract. In this paper, we analyze a non-classical discrete-time queueing model where customers demand variable amounts of work from a server that is able to perform this work at a varying rate. The service demands of the customers are integer numbers of *work units*. They are assumed to be independent and identically distributed (i.i.d.) random variables. The service capacities, i.e., the numbers of work units that the server can process in the consecutive slots, are also assumed to be i.i.d. and their common probability generating function (pgf) is assumed to be rational. New customers arrive in the queueing system according to a general independent arrival process. For this queueing model we present an analysis method, which is based on complex contour integration. Expressions are obtained for the pgfs, the mean values and the tail probabilities of the customer delay and the system content in steady state. The analysis is illustrated by means of some numerical examples.

1. **Introduction.** In many queueing applications, where some kind of customers require some kind of service from a given service facility (or "server"), the amounts of service that the customers demand are different, and the rate at which the server is able to provide that service also varies over time. In some applications, the service rate varies between only 2 possible values, corresponding to "on" and "off" states of the server. This is typically modeled using the concept of vacations or service interruptions (see e.g. [13, 14, 26, 27]). In other applications, where the server is not always either on or off but may also be working at a reduced (or accelerated) speed, the service rate can take on a range of possible values. Example applications where both the service demands and the service rates vary over a range of possible values include packet-switched routers where the packet sizes are variable and the available bandwidth fluctuates over time due to network congestion or the time-varying or fading nature of wireless channels (see e.g. [5, 10, 19, 20]), manufacturing plants where the production capacity varies due to maintenance actions, gradual deterioration or failures of machines (see e.g. [4, 15]), web services where the available processing power fluctuates due to background processes or shared hosting, etc.

2010 *Mathematics Subject Classification.* Primary: 60K25, 90B22; Secondary: 68M20.

Key words and phrases. Queueing theory, discrete time, service demands, service capacities, probability generating functions.

The reviewing process of the paper was handled by Wuyi Yue and Yutaka Takahashi as Guest Editors.

* Corresponding author: Michiel De Muynck.

These variable service requirements and variable service rates are traditionally modeled using the single notion of "service time", which is the amount of time that the server needs to fully serve one customer (see e.g. [2, 17, 30]). This single notion, however, is not always sufficient to model both effects. This is because if the amount of service that each customer requires (which we refer to as the "service demand" of the customer) varies from customer to customer, and the amount of work that the server is able to perform per time unit (which we refer to as the "service capacity" of the server) varies over time, then there may be a non-trivial correlation between the consecutive service times. Indeed, the service time of a customer depends on the service capacity of the server during the service of that customer, which in turn may be correlated with the service capacity during the service of the next customer, which in turn influences the service time of that next customer. The degree of correlation between consecutive service times may depend on many system parameters, including the load and the arrival process in general. Indeed, if the load is very low, then there will most often be long idle times between the service of one customer and the next, so there will likely be little or no correlation between the service times. On the contrary, if the load is high, the amount of correlation may be very large. These kinds of effects cannot be modeled by the many classical queueing models that take the arrival process and the service times to be independent of each other.

In the scientific literature some papers do exist on continuous-time queuing models where the variable service demands of customers and the variable service rates of the server are modeled explicitly (see e.g. [6, 18, 21, 24, 28]). However, the discrete-time equivalents with variable service demands and variable service capacities have received only little attention so far.

In this paper, we therefore study a discrete-time queueing model that explicitly models both the service demands of the customers and the service capacity of the server. Specifically, in our queueing model, time is divided into fixed-length intervals, referred to as (time) slots, and both the service demands of the customers and the service capacities in each time slot are assumed to be integer numbers of "work units". The service demands of the customers are assumed to be independent and identically distributed (i.i.d.) from customer to customer, and the service capacities are i.i.d. from slot to slot. Finally, also the numbers of customer arrivals during the consecutive slots are assumed to be i.i.d. random variables.

With a focus on discrete-time queueing models, existing related work considered various restrictions for the distribution of the service capacities. In [8] and [31], it was assumed that the service capacities follow a geometric distribution. In [9], the model was analyzed under the restriction that the service capacities are deterministically equal to a given constant. In [32], Yao et al. obtained a relationship between the customer-delay and the system-content distributions for the specific case of constant service demands and constant service capacities. In [11] we analyzed this model with the restriction that the distribution of the service capacities has finite support. Finally, in [12], we considered the case where the probability generating function (pgf) of the service capacities is a rational function; the analysis in [12] was, however, restricted to the steady-state customer delay.

In our present paper, we now again consider service capacities with a rational pgf and we extend the analysis of [12] to include the system content as well. Note that all phase-type distributions have a rational pgf (see e.g. [23]), and note in particular that all the restrictions on the service capacities in previous work imply that the

service capacities follow a phase-type distribution. Therefore, our present paper can be seen as a generalization of the papers [8, 31, 9, 11, 12].

The paper is organized as follows. In Section 2, we describe the considered queueing model in more detail. Next, in Section 3, we outline the analysis of the queueing model and present expressions for the pgfs of the unfinished work in the system, the delay of an arbitrary customer, and the system content in steady state. The calculation of moments of the system content and the delay is discussed in Section 4, and approximations for the tail probabilities of the system content and the delay are derived in Section 5. Section 6 is devoted to a remarkable property, referred to as the "invariance property". We discuss some numerical examples in Section 7 and conclusions are given in Section 8.

2. **Queueing model description.** In this paper, we study a discrete-time queueing model, where time is divided into contiguous fixed-length intervals, referred to as (time) slots. New customers arrive in the queueing system according to a general independent arrival process. This means that the numbers of arriving customers during the consecutive slots are i.i.d. from slot to slot. We denote pgf of the number of customers arriving in an arbitrary slot by $A(z)$. The mean number of customers arriving per slot, the so-called mean arrival rate, is denoted by $\lambda \triangleq A'(1)$.

Each customer has a *service demand*, expressed as a positive integer number of work units. This is exactly the amount of work that the server will have to perform, possibly over the course of multiple time slots, to completely serve the customer. The service demands of the customers are assumed to be i.i.d. from customer to customer. The common pgf of the service demands of the customers is denoted by $S(z)$. The mean service demand is denoted by $\tau \triangleq S'(1)$.

The number of work units that the server can execute in a time slot is referred to as the *service capacity* of the server during that time slot. These service capacities are assumed to be non-negative integers that are i.i.d. from slot to slot. We denote their common pgf by $R(z)$, which is assumed to be a rational function. The mean service capacity (per slot) is denoted by $\mu \triangleq R'(1)$. We also introduce the mutually prime polynomials $P_R(z)$ and $Q_R(z)$ such that

$$R(1/z) = \frac{P_R(z)}{Q_R(z)}, \tag{1}$$

and we let m denote the degree of $Q_R(z)$.

We assume that the numbers of arrivals in each slot, the service demands of the customers, and the service capacities during each slot, are mutually independent random variables.

The server cannot initiate the service of a customer during the arrival slot of that customer. Stated otherwise, the service of a customer can start at the earliest during the slot following his arrival slot, even if the customer arrives in an empty system. Customers from the queue are served sequentially by the server in first-come-first-served (FCFS) order. In each slot, no more work units are executed than the available service capacity for that slot. If the available service capacity during a slot is less than the (remaining) service demand of the customer currently in service, then that customer's service simply continues in the next slot, with a reduced remaining service demand. Conversely, if the service capacity is larger than the remaining service demand of the customer currently in service, the server will completely serve that customer and use its remaining service capacity to immediately (i.e., still during the same slot) start also the service of the next customer in the queue (if

any). This is repeated until either the whole service capacity of the server during that slot has been used or there are no customers left in the queue that still require service.

3. Queueing analysis.

In this section, we present the analysis of the above queueing model. Specifically, we derive expressions for the steady-state pgfs of the unfinished work in the system, the customer delay and the system content. Additional details on the analysis method are given in the appendices.

3.1. Pgf of the unfinished work.

As a first step in the analysis, we derive an expression for the pgf $U(z)$ of the unfinished work, i.e., the sum of the (remaining) service demands of all the customers in the system, at the beginning of an arbitrary slot in steady state. This pgf is a useful intermediate result for the delay analysis of Section 3.2.

We begin our derivation by introducing a notation for some of the random variables pertaining to the state of the system in slot k. We let U_k denote the unfinished work at the beginning of slot k, we define A_k as the number of customer arrivals during slot k, the variable $S_{k,i}$ represents the service demand of the ith ($i = 1, 2, ..., A_k$) customer entering the system during slot k and R_k is the service capacity during slot k. In every slot k, the following system equation then holds between these random variables:

$$U_{k+1} = (U_k - R_k)^+ + \sum_{i=1}^{A_k} S_{k,i}, \tag{2}$$

where $(...)^+ = \max(..., 0)$.

Let us now assume that the system is stable, i.e, that the equilibrium condition $\lambda\tau < \mu$ is satisfied. Taking the z-transform of both sides of equation (2), taking the limit for $k \to \infty$ and using the fact that all the above random variables pertaining to the same slot k (i.e., U_k, A_k, R_k and the $S_{k,i}$'s) are all independent of each other, we then easily obtain

$$U(z) = A(S(z)) \lim_{k\to\infty} E\left[z^{(U_k - R_k)^+}\right]. \tag{3}$$

In Appendix A, we show that under the assumption that the service-capacity pgf $R(z)$ is a rational function (see (1)), equation (3) can be further transformed into the following expression for the pgf $U(z)$:

$$U(z) = (\mu - \lambda\tau)\frac{(z-1)A(S(z))}{1 - R(1/z)A(S(z))} \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{z-\zeta}\right)^{\mu_\zeta} \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{z-\xi}{1-\xi}\right)^{n_\xi}, \tag{4}$$

where $\mathcal{S}_R^{-1}$ denotes the set of poles of $R(1/z)$, μ_ξ denotes the multiplicity of a pole $\xi \in \mathcal{S}_R^{-1}$, $\mathcal{N}_T^-$ denotes the set of zeros of $T(z) \triangleq Q_R(z) - A(S(z))P_R(z)$ inside or on the unit circle, excluding the zero at $z = 1$, and n_ξ denotes the multiplicity of a zero $\xi \in \mathcal{N}_T^-$.

3.2. Pgf of the customer delay.

We now turn to the analysis of the customer delay. More specifically, we derive in this section an expression for the pgf $D(z)$ of the delay D_C that an arbitrary customer C experiences in the system in steady state, under a FCFS scheduling discipline. This delay is measured as the number of slots between the end of the arrival slot of the customer and the end of the slot during which the customer leaves. Note that the customer delay cannot be 0, since

a customer that arrives during a slot cannot receive any service during that same slot.

Before proceeding with the actual delay analysis, we first derive the pgf of a related quantity, V_C, the unfinished work observed by the customer C upon arrival. It is defined as the total number of work units present in the system just after the arrival slot of customer C, but to be executed before or during the service of customer C. Work units belonging to any customer arriving after customer C are not counted, even if some of those work units are executed while customer C is technically still in the system, i.e., during the last slot of the service of customer C. Mathematically, V_C is therefore defined as

$$V_C = (U_J - R_J)^+ + \sum_{i=1}^{F_C+1} S_{J,i}, \tag{5}$$

where J denotes the arrival slot of customer C, F_C is the number of customers that arrive in slot J but are to be served before C, and $S_{J,i}$ is the service demand of the ith customer in slot J. It is well-known (see e.g. [25]) that for any queue with independent, ordered arrivals, the pgf of F_C is given by

$$E[z^{F_C}] = \frac{A(z)-1}{\lambda(z-1)}. \tag{6}$$

Using this property and equation (3), the pgf $V(z)$ of V_C then follows immediately as

$$V(z) = \frac{U(z)}{A(S(z))} \cdot \frac{A(S(z))-1}{\lambda\,(S(z)-1)} \cdot S(z) \tag{7}$$

$$= \frac{\mu-\lambda\tau}{\lambda} \cdot \frac{S(z)}{S(z)-1} \cdot \frac{z-1}{1-R(1/z)A(S(z))} \cdot (A(S(z))-1) \cdot \prod_{\zeta\,\in\,\mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{z-\zeta}\right)^{\mu_\zeta} \prod_{\xi\,\in\,\mathcal{N}_T^-} \left(\frac{z-\xi}{1-\xi}\right)^{n_\xi}. \tag{8}$$

The delay D_C of customer C is related to the quantity V_C as follows. Customer C will still be in the system at the start of a slot if and only if fewer than V_C work units have been executed since the end of the arrival slot J of C. In other words,

$$D_C > k \quad \Leftrightarrow \quad V_C > R_{J+1} + R_{J+2} + \ldots + R_{J+k}. \tag{9}$$

In Appendix B, the relationship (9) between random variables D_C and V_C is transformed into the following relationship between their pgfs:

$$D(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V(\xi)}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-zR(1/\xi)}\,\mathrm{d}\xi, \tag{10}$$

where L' is a contour around the origin such that $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|zR(1/\xi)| < 1$, with $\mathcal{R}_X$ denoting the radius of convergence of a pgf $X(z)$.

Next, under the assumption of a rational pgf $R(1/z)$, it is shown in Appendix B that the relation (10) further simplifies to

$$D(z) = \frac{z-1}{z} \sum_{k=0}^{m-1} \frac{V(\alpha_k(z))}{R'(1/\alpha_k(z))} \cdot \frac{\alpha_k(z)}{\alpha_k(z)-1}, \tag{11}$$

where the functions $\alpha_k(z)$ are the m zeros for ξ of

$$1 - zR(1/\xi). \tag{12}$$

The relationship (11) is valid for all z for which the zeros $\alpha_k(z)$ are distinct, which is the case for all but at most $2m-1$ values of z.

Substituting the expression (8) that we previously found for the pgf $V(z)$, and using $R(1/\alpha_k(z)) = 1/z$ (see (12)) to simplify the result, we finally obtain

$$D(z) = \frac{\mu - \lambda\tau}{\lambda} \sum_{k=0}^{m-1} \frac{1-z}{R'(1/\alpha_k(z))} \cdot \frac{S(\alpha_k(z))}{S(\alpha_k(z))-1} \cdot \frac{1-A(S(\alpha_k(z)))}{z-A(S(\alpha_k(z)))} \cdot \alpha_k(z) \cdot \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{\alpha_k(z)-\zeta}\right)^{\mu_\zeta} \cdot \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{\alpha_k(z)-\xi}{1-\xi}\right)^{n_\xi}. \tag{13}$$

This expression still contains the functions $\alpha_k(z)$. These are the zeros of an m-degree polynomial with coefficients that depend on z, and for these zeros a closed-form solution is only available for specific classes of distributions (see Appendix D). If no closed-form solution is available for these zeros, then inverting the pgf $D(z)$ analytically is very difficult. However, inverting this pgf numerically, using methods such as those described in [1], is always very straightforward. Additionally, expression (13) can be used to calculate the expected value and other moments of the delay, as detailed in Section 4, and may also be used to derive a dominant-pole approximation for the tail probabilities of the delay, as explained in Section 5.

3.3. **Pgf of the system content.** The last quantity that we study in this paper is the system content. Specifically, in this section, we obtain an expression for the pgf $B(z)$ of the system content, i.e., the total number of customers present in the queueing system, at the beginning of an arbitrary slot in steady state.

To do so, we first relate the distribution of the system content at the beginning of an arbitrary slot to the system-content distributions at arrival and departure instants. It is a well-known fact (see e.g. [22]) that in any system with ordered arrivals and departures, the pgf $B_a(z)$ of the system content just before the arrival of an arbitrary customer is equal to the pgf $B_d(z)$ of the system content just after the departure of an arbitrary customer, i.e., $B_a(z) = B_d(z)$. To relate $B(z)$ and $B_a(z)$, we simply use (6) and find

$$B_a(z) = B(z)E[z^{F_C}] = B(z)\frac{A(z)-1}{\lambda(z-1)}. \tag{14}$$

Secondly, we relate the system content just after the departure of an arbitrary customer C to the delay D_C of that customer. Since customers are served in FCFS order, the system content just after the departure of C is given by

$$G_C + \sum_{k=1}^{D_C} A_{C,k}, \tag{15}$$

where G_C is the number of customers arriving in the same slot as customer C, but to be served after C, and $A_{C,k}$ is the number of customers arriving in the kth slot after the arrival slot of customer C. Note that G_C and D_C are conditionally independent when given F_C. Also note that similar to (6), the joint distribution of G_C and F_C can be related to the distribution of the number of customers A arriving in an arbitrary slot in steady state, as follows:

$$\text{Prob}[G_C = g, F_C = f] = \frac{1}{\lambda}\,\text{Prob}[A = g + f + 1]. \tag{16}$$

Therefore, we can derive the pgf $B_d(z)$ of the system content at departure instants as

$$\begin{aligned} B_d(z) &= E[z^{G_C+\sum_{k=1}^{D_C} A_{C,k}}] \\ &= \sum_{f=0}^{\infty} \text{Prob}[F_C = f] \cdot E[z^{G_C}|F_C = f] \cdot E[z^{\sum_{k=1}^{D_C} A_{C,k}}|F_C = f]. \\ &= \sum_{f=0}^{\infty} \frac{D_f(A(z))}{\lambda z^{f+1}} \sum_{i=f+1}^{\infty} \text{Prob}[A = i] z^i, \end{aligned} \tag{17}$$

where $D_f(z)$ denotes the conditional pgf of the delay of an arbitrary customer C, given that F_C takes the integer value f.

Next, using similar methods as before, the conditional pgf $D_f(z)$ and, in view of the intermediate results (14) and (17), also the pgf $B(z)$ can be related to the pgf $U(z)$ of the unfinished work at the beginning of an arbitrary slot. In particular, in Appendix C, the following relationship is obtained:

$$\begin{aligned} B(z) = \frac{A(z)(z-1)}{2\pi\imath(A(z)-1)} \oint_{L'} \frac{U(\xi)}{A(S(\xi))} \cdot \frac{1}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-A(z)R(1/\xi)} \\ \cdot S(\xi) \cdot \frac{A(z)-A(S(\xi))}{z-S(\xi)}\,\mathrm{d}\xi, \end{aligned} \tag{18}$$

where L' is a contour around the origin such that $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|A(z)R(1/\xi)| < 1$.

Moreover, it is shown in Appendix C that for a rational pgf $R(z)$ of the service capacities, the contour integral in this expression can be further simplified to

$$B(z) = \frac{z-1}{A(z)} \sum_{k=0}^{m-1} \frac{U(\beta_k(z))}{R'(1/\beta_k(z))} \frac{\beta_k(z)}{\beta_k(z)-1} \frac{A(z)-A(S(\beta_k(z)))}{z-S(\beta_k(z))} \frac{S(\beta_k(z))}{A(S(\beta_k(z)))}, \tag{19}$$

where the functions $\beta_k(z)$ are defined as $\alpha_k(A(z))$, $k = 0, 1, ..., m-1$, i.e., the zeros for ξ of

$$1 - A(z)R(1/\xi). \tag{20}$$

The relationship (19) is valid for all z for which these zeros $\beta_k(z)$ are distinct, which is the case for all z but an isolated set; see Appendix D for further discussion on these functions $\beta_k(z)$.

Finally, combining equation (19) with the result (4) for $U(z)$, we obtain the following expression for $B(z)$:

$$\begin{aligned} B(z) = (\mu - \lambda\tau)(z-1) \sum_{k=0}^{m-1} \frac{\beta_k(z)}{R'(1/\beta_k(z))} \cdot \frac{S(\beta_k(z))}{z-S(\beta_k(z))} \\ \cdot \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{\beta_k(z)-\zeta}\right)^{\mu_\zeta} \cdot \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{\beta_k(z)-\xi}{1-\xi}\right)^{n_\xi}. \end{aligned} \tag{21}$$

4. **Moments of the system content and the customer delay.** In this section, we derive explicit expressions for the first-order moments, i.e., the expected values, of the system content at the beginning of an arbitrary slot and the delay of an arbitrary customer in steady state, and we explain how to obtain expressions for higher-order moments of these quantities.

An expression for the mean system content can be obtained by evaluating the first derivative of $B(z)$ to z at $z = 1$. This is however not entirely trivial, as it requires the determination of the derivatives of the implicitly defined functions $\beta_k(z)$ at $z = 1$. These latter derivatives can be calculated by differentiating both sides of equation (20), which yields

$$\beta_k'(z) = \frac{A'(z)\,\beta_k(z)^2}{A(z)^2\,R'(1/\beta_k(z))}. \tag{22}$$

Higher-order derivatives of $\beta_k(z)$ may be obtained by differentiating both sides of equation (22) repeatedly.

Evaluation of the derivative of $B(z)$ at $z = 1$ also requires the use of l'Hôpital's rule, since it can be seen from equation (20) that one of the zeros $\beta_k(z)$ equals 1 when $z = 1$. This zero must have multiplicity 1, since $R'(1) > 0$, so we may unambiguously denote this zero as $\beta_0(z)$, so that $\beta_0(1) = 1$. Under a few assumptions which we will discuss shortly, we may now obtain an expression for the mean system content by differentiating (21) to z, substituting $z = 1$, and using l'Hôpital's rule on the term for $k = 0$. This yields

$$\begin{aligned} B'(1) = \frac{\lambda}{\mu}\Bigg(1 + \tau - \sum_{\zeta\,\in\,\mathcal{S}_R^{-1}} \frac{\mu_\zeta}{1-\zeta} + \sum_{\xi\,\in\,\mathcal{N}_T^-} \frac{n_\xi}{1-\xi}\Bigg) \\ + \frac{R''(1)\lambda(2\mu - \lambda\tau) + S''(1)\lambda^2\mu + A''(1)\mu^2\tau + 2\tau\lambda^2\mu(1-\mu)}{2(\mu-\lambda\tau)\mu^2} \\ + (\mu - \lambda\tau)\sum_{k=1}^{m-1} \frac{\beta_k(1)}{R'(1/\beta_k(1))} \cdot \frac{S(\beta_k(1))}{1 - S(\beta_k(1))} \\ \cdot \prod_{\zeta\,\in\,\mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{\beta_k(1)-\zeta}\right)^{\mu_\zeta} \cdot \prod_{\xi\,\in\,\mathcal{N}_T^-} \left(\frac{\beta_k(1)-\xi}{1-\xi}\right)^{n_\xi}. \end{aligned} \tag{23}$$

The assumptions that must hold are the following:

- All zeros $\beta_k(1)$ must be distinct. If they are not, then equation (21) is not applicable for $z = 1$. The mean value of the system content may then be obtained by evaluating the derivative with respect to z at $z = 1$ of both sides of equation (72) instead. However, since this is relatively difficult in general, it may be easier to approximate $B'(1)$ numerically by using a finite difference method in that case.
- The term for $k = 0$ must be the only term of equation (21) that requires l'Hôpitals rule for the evaluation of the derivative at $z = 1$. It can be shown that the assumption does not hold if and only if the greatest common divisor g of the period of the service demands and the period of the service capacities is greater than 1. In that case, the simple division of all service demands and capacities by g yields a queueing system that behaves exactly the same way, and in particular it has the same moments of the system content and the delay. For this equivalent queueing system, g will be equal to 1, so the expression (23) (applied to the equivalent queueing system) may be used to calculate the mean system content.

The mean value $D'(1)$ of the customer delay may be obtained similarly by evaluating the derivative of equation (13) at $z = 1$. The obtained expression is the same as (23) multiplied by a factor $1/\lambda$, in agreement with Little's law.

By using the moment-generating property of probability generating functions, also higher-order moments of the system content and delay may be obtained. This requires the evaluation of higher-order derivatives of $B(z)$ or $D(z)$ at $z = 1$. The calculations, e.g. of the variances of the system content and the delay, are rather tedious and the resulting expressions are too long to be included here. We do, however, show some numerical results in Section 7.

5. **Tail probabilities of the system content and the customer delay.** In this section, we describe dominant-pole approximations for the tail probabilities of the system content and the customer delay in steady state. As explained in [7], if z_X is the smallest positive real-valued pole of the pgf $X(z)$ of a random variable X, then the tail probabilities of X can be approximated as

$$\text{Prob}[X = n] \approx -C_X z_X^{-n-1}, \tag{24}$$

and

$$\text{Prob}[X > n] \approx \frac{-C_X}{z_X(z_X - 1)} z_X^{-n}, \tag{25}$$

where C_X is the residue of $X(z)$ at $z = z_X$. These approximations thus require the determination of 2 constants: z_X and C_X. We will describe how to determine these constants, first for the system content and then for the customer delay.

5.1. **System content.** In general, the dominant pole z_B of the system content has to be found numerically. One way of doing this is by finding the smallest positive real-valued zero of $B(1/z)$. Since this is a real-valued zero, many methods can be used for this, such as the bisection method or the Illinois algorithm.

If it is known a priori that the zeros $\beta_k(z_B)$ are distinct, e.g. if the service-capacity distribution belongs to one of the classes of distributions described in Appendix D, then it is not necessary to use the lengthy expressions for $B(z)$ obtained in Section 3.3. Instead, the following theorem is available for a much faster determination of z_B:

Theorem 5.1. *If the zeros $\beta_k(z_B)$ are distinct, then the dominant pole z_B of $B(z)$ is the smallest positive real-valued zero from all the zeros of $z - S(\beta_k(z))$, $k = 0, 1, ..., m-1$.*

Proof. From expression (21) for $B(z)$ it is clear that z_B must be one of the following:

(a) A zero of $\beta_k(z) - \zeta$ for some $k \in [0, m-1]$ and some $\zeta \in \mathcal{S}_R^{-1}$. However, note that this would imply that $R(1/\beta_k(z_B)) = R(1/\zeta)$, but by (20), it follows that $R(1/\beta_k(z_B)) = 1/A(z_B)$, while by definition of $\mathcal{S}_R^{-1}$, we have that $R(1/\zeta) = \infty$. Therefore, $A(z_B)$ would have to be 0. But by (14), (17) and the fact that $D_f(0) = 0$, $B(z_B)$ would then have to be 0 as well, but z_B is supposed to be a pole of $B(z)$. From this contradiction we conclude that $\beta_k(z_B) \neq \zeta$ for all k and ζ.

(b) A pole of $\beta_k(z)$ for some k. However, this would not lead to a pole of $B(z)$ because the factors $\beta_k(z)$ and $\beta_k(z) - \xi$ in the numerator have combined multiplicity m, while the factors $\beta_k(z) - \zeta$ in the denominator also have combined multiplicity m.

(c) A zero of $R'(1/\beta_k(z))$ for some k. However, this is not possible due to our assumption that the zeros $\beta_k(z_B)$ are distinct (see (20) and Appendix D).

(d) A pole of $S(\beta_k(z))/(z - S(\beta_k(z))$ for some k. This can only occur if z is a zero of $z - S(\beta_k(z))$.

Since (d) is the only possibility, we conclude that z_B is a zero of $z - S(\beta_k(z))$ for some $k \in [0, m-1]$. Furthermore, since z_B is the dominant pole, it must be the smallest positive real-valued z for which $z = S(\beta_k(z))$ for some $k \in [0, m-1]$. □

Once z_B is known, C_B can be calculated. We again assume that the zeros $\beta_k(z_B)$ are distinct. It is easy to check numerically which terms in the summation over k in (21) contribute to the pole of $B(z)$ at $z = z_B$, i.e., for which values of k $z_B = S(\beta_k(z_B))$. We denote the set of these k as $\mathcal{K}$. The residue C_B of $B(z)$ at $z = z_B$ can now be calculated by summing the individual residues of these terms in (21). We find

$$C_B = (\mu - \lambda\tau) \sum_{k \in \mathcal{K}} \frac{\beta_k(z_B)\,(z_B - 1)\,z_B\,A(z_B)^2}{R'(1/\beta_k(z_B))\,A(z_B)^2 - S'(\beta_k(z_B))A'(z_B)\beta_k(z_B)^2} \cdot \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{\beta_k(z_B) - \zeta} \right)^{\mu_\zeta} \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{\beta_k(z_B) - \xi}{1 - \xi} \right)^{n_\xi}, \tag{26}$$

where we also used the fact that if $k \in \mathcal{K}$, then $S(\beta_k(z_B)) = z_B$ to simplify the result.

5.2. **Customer delay.** For the delay, the dominant pole z_D and the residue C_D can be found as follows. Determining z_D could in principle be done similarly to determining z_B by finding the smallest positive real-valued zero of $1/D(z)$. However, in Appendix B we showed that the radius of convergence $\mathcal{R}_D$ of $D(z)$ is given by $\mathcal{R}_D = 1/R(1/\mathcal{R}_V)$. This implies that $z_D = 1/R(1/z_V)$, where z_V is the dominant pole of $V(z)$. From (7) and (4) it is easy to see that the dominant pole z_V of $V(z)$ is given by the smallest positive real-valued zero of $1 - R(1/z)A(S(z))$. Since this does not contain any implicitly defined functions, finding z_V and using it to calculate z_D is typically easier than finding z_D directly.

To determine the residue C_D, we will use the relation (11). Note that the factor $\alpha_k(z) - 1$ in the denominator of (11) does not yield a pole of $D(z)$ at $z = z_D$ since that would imply $z_D = 1$. Then, using arguments similar to those used in Theorem 5.1, we find that if the zeros $\alpha_k(z_D)$ are distinct, then the factor $R'(1/\alpha_k(z))$ does not contribute to the dominant pole either, so that $\alpha_k(z_D)$ must be a pole of $V(z)$ for some k. Denote the set of all z that are both a pole of $V(z)$ and equal to $\alpha_k(z_D)$ for some $k \in [0, m-1]$ as $\mathcal{V}$. Then, we can calculate the residue C_D as

$$C_D = \frac{\mu - \lambda\tau}{\lambda} \cdot \sum_{z_* \in \mathcal{V}} \frac{S(z_*)}{S(z_*) - 1} \cdot \frac{(z_D - 1)^2}{z_* \left[R'(1/z_*) z_*^{-2} - A'(S(z_*))S'(z_*) {z_D}^{-2} \right]} \cdot \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{z_* - \zeta} \right)^{\mu_\zeta} \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{z_* - \xi}{1 - \xi} \right)^{n_\xi}, \tag{27}$$

where we have used the facts that if $z_* \in \mathcal{V}$, then $\alpha_k(z_D) = z_*$ and $A(S(z_*)) = 1/R(1/z_*) = 1/R(1/\alpha_k(z_D)) = z_D$ to simplify the result.

6. **Invariance property.** In [8] it was first observed that when the service demands follow a shifted geometric distribution (with minimum value 1) and the service capacities follow a geometric distribution, i.e., when

$$S(z) = \frac{z}{\tau + (1-\tau)z} \tag{28}$$

and

$$R(z) = \frac{1}{1+\mu-\mu z}, \tag{29}$$

then the distributions of the system content and the delay depend on τ and μ only through their ratio τ/μ. This remarkable property was termed the "geometric invariance property" in [8]. From numerical results obtained for the more general queueing model considered in this paper (see also further in Section 7), we discovered that this may also be true for some other service-demand and service-capacity distributions. In this section, we therefore first explore other conditions under which the "invariance property" holds, i.e., under which the distributions of the system content and the delay only depend on τ and μ through their ratio τ/μ.

If the service demands follow a shifted geometric distribution, then due to the memoryless property of the geometric distribution, each work unit of a customer's service demand has the same probability $1/\tau$ of being the last work unit of that customer's demand, regardless of how many other work units of service that customer has already received. Therefore, if the system has a certain service capacity of n work units in a given slot, then the number of customers that can leave in that slot is binomially distributed with parameters n and $1/\tau$. Hence, the pgf $\tilde{R}(z)$ of the number of customers that can leave in an arbitrary slot k can be calculated as

$$\begin{aligned}\tilde{R}(z) &= \sum_{n=0}^{\infty} \text{Prob}[R_k = n] \cdot \left(1 - \frac{1}{\tau} + \frac{z}{\tau}\right)^n \\ &= R\left(1 - \frac{1}{\tau} + \frac{z}{\tau}\right).\end{aligned} \tag{30}$$

To determine when the invariance property holds, let c denote the ratio τ/μ. Then (30) can be rewritten as

$$\tilde{R}(z) = R\left(\frac{c\mu - 1 + z}{c\mu}\right). \tag{31}$$

Since the pgf $\tilde{R}(z)$ describes the number of customers that can be served each slot, which is independent from slot to slot, this pgf fully characterizes the service process. Therefore, if the service-capacity distribution is such that (31) only depends on c, and not on the individual value of μ, then the invariance property holds. This is at least the case for the following distributions:

1. The negative binomial distribution, with pgf
$$R(z) = \left(\frac{m}{m+\mu-\mu z}\right)^m, \tag{32}$$
with as special case the geometric distribution ($m = 1$).
2. The binomial distribution, with pgf
$$R(z) = \left(1 - \frac{\mu}{m} + \frac{\mu}{m} z\right)^m, \tag{33}$$
with as special case the Bernoulli distribution ($m = 1$).
3. The Poisson distribution, with pgf $R(z) = \exp(\mu(z-1))$.

Next, if the service capacities follow a geometric distribution, then again due to the memoryless property of the geometric distribution, regardless of how many work units of service have already been executed during a slot, there will be service capacity remaining in that slot with probability $\mu/(1+\mu)$ and no remaining capacity with probability $1/(1+\mu)$. Therefore, the number of slots it takes to serve each

work unit of a customer's service demand simply follows a geometric distribution with parameter $1/(1+\mu)$ and mean $1/\mu$. Hence, if a customer has a certain service demand n, then his "service time" is negative binomially distributed with parameters $1/(1+\mu)$ and n and mean n/μ. Note that this service time is independent from one customer to the next, and can be equal to 0 slots. This allows us to calculate the pgf $\tilde{S}(z)$ of the service time of an arbitrary customer C as

$$\begin{aligned}\tilde{S}(z) &= \sum_{n=0}^{\infty} \text{Prob}[S_C = n] \cdot \left(\frac{\mu}{1+\mu-z}\right)^n \\ &= S\left(\frac{\mu}{1+\mu-z}\right).\end{aligned} \tag{34}$$

Again denoting the fraction τ/μ by c, we rewrite this as

$$\tilde{S}(z) = S\left(\frac{\tau}{c+\tau-cz}\right). \tag{35}$$

Similarly as before, the pgf $\tilde{S}(z)$ fully characterizes the service process, so that if the service-demand distribution is such that expression (35) does not depend on τ but only on c, then the behavior of the queue only depends on the value of τ (or μ) through the ratio τ/μ, i.e., the invariance property holds. This is for instance the case for the shifted negative binomial distribution, with pgf

$$S(z) = \left(\frac{mz}{\tau+(m-\tau)z}\right)^m, \tag{36}$$

with as special case the shifted geometric distribution ($m = 1$).

For the above combinations of service-demand and service-capacity distributions we know for sure that the invariance property holds. There may, however, be other combinations for which it holds, and in particular, there may be combinations for which the invariance property holds where neither the service demands nor the service capacities are geometrically distributed. The invariance property, however, does not hold in general. This will be illustrated in the next section, where numerical results (see Figure 3) indicate that the property does not hold anymore for the combination of shifted geometric demands and shifted geometric or deterministic service capacities. A full classification of all the conditions under which the invariance property holds thus remains an open question.

7. **Numerical examples and discussion.** In this section, we present a few numerical examples that illustrate the behavior of the queueing system, and in particular the impact of the service-capacity distribution on several performance measures of the system. Throughout this section, we consider Poisson arrivals with mean arrival rate λ, i.e., $A(z) = e^{\lambda(z-1)}$.

In the first example, shown in Figures 1 and 2, we study the impact of the service-capacity distribution on the mean and the variance of the customer delay under varying loads $\rho = \lambda\tau/\mu$. Here, the load ρ is varied by varying λ, the service demands are constant, with $\tau = 11$, and 4 different service-capacity distributions are considered, all with mean $\mu = 10$: deterministic capacities with pgf

$$R(z) = z^{10}, \tag{37}$$

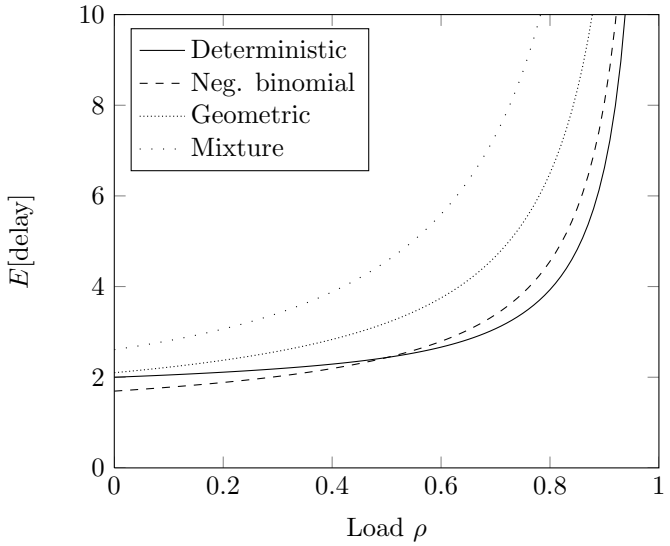

FIGURE 1. Mean customer delay versus the load ρ, for Poisson arrivals with varying λ, deterministic service demands with $\tau = 11$ and various service-capacity distributions (as indicated), all with mean $\mu = 10$.

negative binomial capacities with parameter $m = 5$ and pgf

$$R(z) = \frac{1}{(3-2z)^5}, \tag{38}$$

geometric capacities with pgf

$$R(z) = \frac{1}{11-10z}, \tag{39}$$

and a weighted mixture of 2 geometric distributions with means 5 and 30 such that the overall mean $\mu = 10$, with corresponding pgf

$$R(z) = \frac{26-25z}{(6-5z)(31-30z)}.$$

The variances of these 4 distributions are respectively 0, 30, 110, and 310.

From Figure 1 it can be seen that, generally, a higher variance of the service capacity leads to a higher mean delay. This is also what one would expect intuitively, since more variability on the service capacities is in general expected to lead to a burstier service process, which should in turn cause longer queues. However, we notice in Figure 1 also one case where the opposite is true: under low load, the system with negative binomially distributed service capacities turns out to have a lower mean delay than the system with deterministic capacities, despite the fact that the deterministic distribution has the lowest variance. Under high load it is the other way around again.

This observation can be explained as follows. When the load ρ is very close to 0, almost every customer arrives in an empty system. Due to the fact that the service demands are deterministically equal to 11 work units, any customer arriving

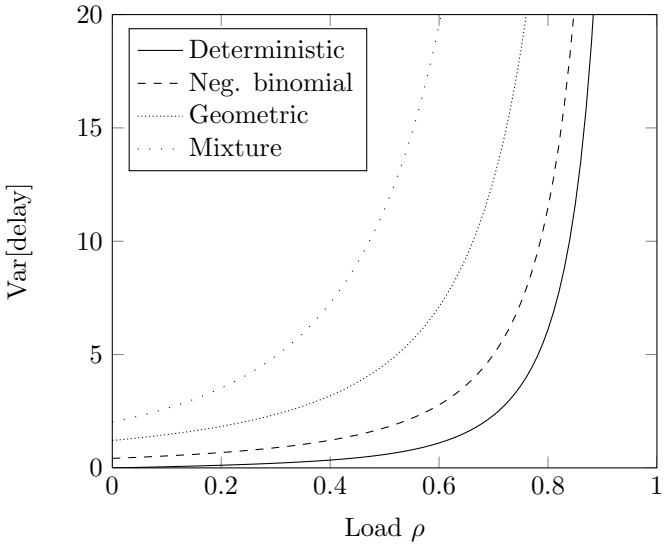

FIGURE 2. Variance of the customer delay versus the load ρ, for Poisson arrivals with varying λ, deterministic service demands with $\tau = 11$ and various service-capacity distributions (as indicated), all with mean $\mu = 10$.

in an empty system will be served in one slot if and only if the service capacity in the next slot is at least 11 work units. Since this is never the case for deterministic service capacities (with $\mu = 10$), the minimum delay in this case is 2 slots. Indeed, in Figure 1 it can be seen that for deterministic service capacities, the mean delay goes to 2 as ρ goes to 0. For negative binomial service capacities however, the probability that the service capacity is at least 11 work units in a slot is approximately 40.4%, so that around 40.4% of the customers arriving in an empty system will be served in one slot, and if the load is low enough then the mean delay will be lower than 2 slots, as can be observed in Figure 1. On the contrary, when the load is high, relatively few customers arrive in an empty system, so the benefit that negative binomial service capacities give to these customers becomes insignificant. In this case, the system with deterministic service capacities will outperform the system with negative binomial service capacities, due to the lack of randomness in the service process, which allows customers to be served more regularly.

From the above discussion, we conclude that the queueing model studied in this paper cannot be reduced to a classical model with independent service times, because the "mean service time" depends on the arrival process and on the load ρ in particular. For instance, for deterministic service capacities, this mean service time approaches 2 as ρ approaches 0, whereas it approaches 11/10 as ρ approaches 1.

In Figure 2, we show the variance of the customer delay for the same system parameters as in Figure 1. We observe that the relative ordering of the variances of the customer delay for the various service-capacity distributions remains the same for all values of the load ρ, i.e., the lines in Figure 2 do not cross, unlike those in Figure 1. In Figure 2, a higher variance of the service capacity always corresponds to a higher variance of the customer delay. In particular, the variance of the customer

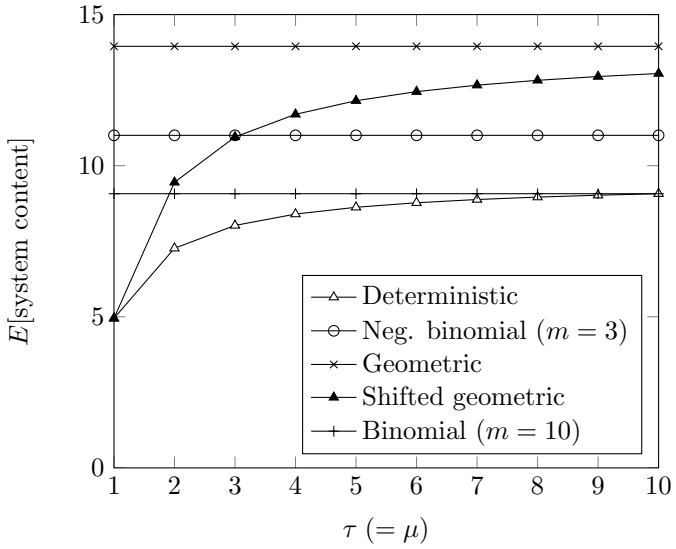

FIGURE 3. Mean system content versus the mean service demand τ for Poisson arrivals with $\lambda = 0.9$, shifted geometric service demands and various service-capacity distributions (as indicated), with mean $\mu = \tau$.

delay is lower for deterministic service capacities than for negative binomial service capacities for all values of ρ, including those near 0. In fact, as ρ approaches 0, then for deterministic service capacities, the variance of the customer delay approaches 0 as well, which is not the case for negative binomial service capacities. This is because if the load is very low, then with deterministic service capacities almost all customers have a delay of 2 slots, whereas with negative binomial service capacities, around 40.4% of the customers have a delay of 1 slot, the rest has a delay of 2 or more slots, as discussed earlier.

In a second example, shown in Figure 3, we keep the load ρ and the mean arrival rate λ fixed at $\rho = \lambda = 0.9$, so we keep the ratio $\tau/\mu = 1$ and we scale τ and μ together to observe the impact of their actual values on the mean system content. The service demands have a shifted geometric distribution (with minimum value 1) and 5 different service-capacity distributions are considered, all with mean $\mu = \tau$: deterministic, negative binomial (with $m = 3$), geometric, shifted geometric (with minimum value 1), and binomial (with $m = 10$) service capacities.

We clearly see that for geometric service capacities (and shifted geometric demands), the mean system content does not depend on the actual values of τ and μ but merely on their ratio, as predicted by the invariance property discussed in Section 6. This means that higher service demands of the customers are in this case exactly compensated by proportionally equally higher service capacities of the system. However, note that while the invariance property holds for geometric service capacities, it does not hold for *shifted* geometric service capacities, as the mean system content in the latter case clearly depends on μ. The same is true for deterministic service capacities. Finally, from Figure 3 it can also be seen that the invariance property also holds for binomial and negative binomial service capacities, in agreement with the discussion in Section 6.

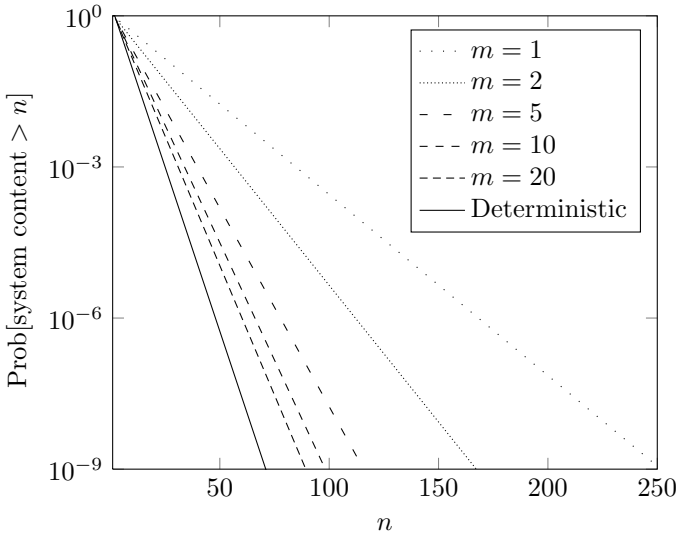

FIGURE 4. Dominant-pole approximation of the tail probabilities of the system content, for Poisson arrivals with $\lambda = 3$, uniformly distributed service demands from 1 to 10 work units, and negative binomial service capacities with $\mu = 10$ and various values of the parameter m, as well as deterministic service capacities.

In our final numerical example, shown in Figure 4, we consider Poisson arrivals with $\lambda = 3$ and service demands that are uniformly distributed between 1 and 10 work units. The service capacities follow a negative binomial distribution with $\mu = 10$ and parameter m, and we examine the influence of the parameter m on the tail probabilities of the system content. We observe from Figure 4 that an increase in the parameter m reduces the probability that the system content becomes large. This could be expected intuitively, since an increase of m corresponds to a decrease of the variance of the service capacities, which makes the system serve the customers at a more regular rate, which in turn decreases the probability of a large system content.

We moreover see that the impact of increasing m is the largest when m is very small. For instance, the increase from $m = 1$ to $m = 2$ corresponds to a relatively large performance gain, whereas the step from $m = 10$ to $m = 20$ results in a comparatively small performance gain. This is because there is an upper limit to the amount of performance that can be gained by increasing m further and further. In Figure 4, we also show the tail probabilities for the case of deterministic service capacities, which can be considered as negative binomial service capacities with parameter $m = \infty$. The full line in Figure 4 therefore represents a lower bound on the tail probabilities that can be achieved by changing the parameter m.

8. **Conclusion.** In this paper, we analyzed a discrete-time queueing model with general service demands and service capacities with rational pgf. Our main results are the obtained analytical expressions for the pgfs of the steady-state customer delay and the system content, the expressions for the mean values of the system

content and the delay, and the dominant-pole approximations for the tail probabilities of the system content and the delay. We also found new sufficient conditions under which the invariance property holds.

While the considered model is very general, allowing arbitrary distributions of the number of arrivals per slot and the service demands of the customers, and arbitrary phase-type distributions for the service capacity per slot, a restriction of the model is that the service capacities are assumed to be independent from slot to slot. Studying the effects of correlation between service capacities is a compelling direction for future research.

Acknowledgments. This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

Appendix A. **Derivation of the pgf $U(z)$.** In this appendix, we derive the expression (4) for the pgf $U(z)$ of the unfinished work at the beginning of an arbitrary slot in steady state. Since the random variables U_k and R_k are independent and R_k is a random variable with a rational generating function, we can use a method based on complex contour integration, similar to the one presented in [29] (for the analysis of the classical discrete-time $\mathrm{G}^{(\mathrm{G})}/\mathrm{Geo}/1$ queue), to further work out the above equation. Under the assumption of a stable system, i.e., under the equilibrium condition $\lambda\tau < \mu$, the following equation is then obtained for the steady-state pgf $U(z)$ of the unfinished work at the beginning of a slot (see [29]):

$$U(z) = A(S(z)) \left[U(z)R(1/z) \, + \, (z-1) \sum_{\zeta \, \in \, \mathcal{S}_R^{-1}} F_\zeta(z) \right]. \tag{40}$$

Equation (40) is valid at least for all z inside the unit circle, with $z \notin \mathcal{S}_R^{-1}$, where $\mathcal{S}_R^{-1}$ denotes the set of singularities of $R(1/z)$. The function $F_\zeta(z)$ in (40) is defined as

$$F_\zeta(z) = \frac{1}{2\pi\imath} \oint_{C_\zeta} \frac{U(\xi)R(1/\xi)}{(\xi - z)(\xi - 1)} \, \mathrm{d}\xi, \tag{41}$$

with $\imath^2 = -1$ and C_ζ a small (counterclockwise) contour around ζ but not around any other singularity of $R(1/\xi)$, nor any singularity of $U(\xi)$, nor around 1 or z.

Let us now assume that the service-capacity pgf $R(z)$ is a rational function. Then all singularities ζ of $R(1/z)$ are poles and we can write

$$R(1/z) = \frac{P_R(z)}{Q_R(z)} = \frac{P_R(z)}{\prod_{\zeta \, \in \, \mathcal{S}_R^{-1}} (z - \zeta)^{\mu_\zeta}}, \tag{42}$$

wherein $P_R(z)$ and $Q_R(z)$ are two mutually prime polynomials and μ_ζ denotes the multiplicity of the singularity $\zeta \in \mathcal{S}_R^{-1}$. Note that the degree of $P_R(z)$ cannot be higher than the degree $m = \sum_{\zeta \, \in \, \mathcal{S}_R^{-1}} \mu_\zeta$ of $Q_R(z)$, since $\lim_{z \to \infty} R(1/z) = R(0) \in [0, 1]$. Therefore, using the expression for the residue of a complex function at a pole ζ with multiplicity μ_ζ, we easily find that the contour integral $F_\zeta(z)$ takes the form

$$F_\zeta(z) = \sum_{k=1}^{\mu_\zeta} \frac{c_k}{(z - \zeta)^k}, \tag{43}$$

for yet unknown constants c_k, which in turn leads to

$$\sum_{\zeta \in \mathcal{S}_R^{-1}} F_\zeta(z) = \frac{N(z)}{Q_R(z)}, \tag{44}$$

with $N(z)$ an unknown polynomial of degree $m-1$. Hence, we get

$$U(z) = \frac{(z-1)A(S(z))N(z)}{Q_R(z) - A(S(z))P_R(z)}. \tag{45}$$

Using Rouché's theorem it can now be shown (see e.g. [3]) that the denominator $T(z) = Q_R(z) - A(S(z))P_R(z)$ has exactly m zeros inside or on the unit circle, one of which is equal to 1. Since $U(z)$ must remain bounded in these zeros, the numerator of $U(z)$ has to vanish as well, with at least the same multiplicity. This completely determines the polynomial $N(z)$ and the pgf $U(z)$ except for a constant factor. With the normalization condition $U(1) = 1$, we finally get the following expression for $U(z)$:

$$U(z) = (\mu - \lambda\tau)\frac{(z-1)A(S(z))}{1 - R(1/z)A(S(z))} \prod_{\zeta \in \mathcal{S}_R^{-1}} \left(\frac{1-\zeta}{z-\zeta}\right)^{\mu_\zeta} \prod_{\xi \in \mathcal{N}_T^-} \left(\frac{z-\xi}{1-\xi}\right)^{n_\xi}, \tag{46}$$

where $\mathcal{N}_T^-$ denotes the set of zeros of $T(z)$ inside or on the unit circle, excluding the zero at $z = 1$, and n_ξ denotes the multiplicity of a zero ξ in this set.

Appendix B. **Relationship between the pgfs $D(z)$ and $V(z)$.** In this appendix, we first prove the following general relationship between the steady-state pgf $D(z)$ of the delay of an arbitrary customer and the steady-state pgf $V(z)$ of the unfinished work observed by an arbitrary customer:

$$D(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V(\xi)}{\xi\,(\xi - 1)} \cdot \frac{1 - R(1/\xi)}{1 - zR(1/\xi)}\,\mathrm{d}\xi, \tag{47}$$

where L' is a contour around the origin such that $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|zR(1/\xi)| < 1$, with $\mathcal{R}_X$ denoting the radius of convergence of a pgf $X(z)$. The radius of convergence $\mathcal{R}_D$ of $D(z)$ is given by $\mathcal{R}_D = 1/R(1/\mathcal{R}_V)$.

Proof. The delay D_C of an arbitrary customer C and the unfinished work V_C observed by a customer C upon arrival are related as follows (see (9)):

$$D_C > k \quad \Leftrightarrow \quad V_C > R_J^{(k)}, \tag{48}$$

where $R_J^{(k)}$ denotes the sum of the service capacities during the k slots following the arrival slot J of customer C. Since these service capacities are independent, the pgf of $R_J^{(k)}$ is given by $R(z)^k$. Using the fact that V_C and $R_J^{(k)}$ are independent, we then find

$$\frac{D(z)-1}{z-1} = \sum_{k=0}^{\infty} \mathrm{Prob}[D_C > k] z^k = \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{i-1} \mathrm{Prob}[V_C = i]\,\mathrm{Prob}[R_J^{(k)} = j] z^k. \tag{49}$$

Next, the inversion formula for probability generating functions states that

$$\mathrm{Prob}[R_J^{(k)} = j] = \frac{1}{2\pi\imath} \oint_L \frac{R(\zeta)^k}{\zeta^{j+1}}\,\mathrm{d}\zeta, \tag{50}$$

where L is a contour around the origin such that $\forall \zeta \in L : |\zeta| < \mathcal{R}_R$, where $\mathcal{R}_X$ denotes the radius of convergence of a pgf $X(z)$. Note that the radius of convergence

of $R(z)$ and that of $R(z)^k$ are equal. By means of (50), we can rewrite equation (49) as

$$\begin{aligned}\frac{D(z)-1}{z-1} &= \frac{1}{2\pi\imath}\sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\sum_{j=0}^{i-1}\oint_L \text{Prob}[V_C=i]\frac{\big(zR(\zeta)\big)^k}{\zeta^{j+1}}\,\mathrm{d}\zeta \\ &= \frac{1}{2\pi\imath}\sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\oint_L \text{Prob}[V_C=i]\,\big(zR(\zeta)\big)^k\,\frac{\zeta^{-i}-1}{1-\zeta}\,\mathrm{d}\zeta.\end{aligned} \tag{51}$$

The above infinite summation of contour integrals is equal to the contour integral of the infinite series (i.e., we may "swap" the summation and integration symbols) if the contour L is chosen such that the resulting infinite series is uniformly convergent. It is important to question when such a contour can be constructed. This is the case if $\forall\zeta\in L : |1/\zeta| < \mathcal{R}_V$ and $|zR(\zeta)|<1$. The former condition imposes a lower bound on $|\zeta|$, whereas the latter imposes an upper bound on $|R(\zeta)|$ that depends on z. Since this upper bound is most severe when ζ is real and positive, and since $R(\zeta)$ is an increasing function of ζ on the part of the real axis where $0\le\zeta<\mathcal{R}_R$, the bounds can be rewritten as $R(1/\mathcal{R}_V) < R(|\zeta|) < |1/z|$. We conclude that a contour can be constructed if and only if $|z| < 1/R(1/\mathcal{R}_V)$. It follows that the radius of convergence $\mathcal{R}_D$ of $D(z)$ is given by $\mathcal{R}_D = 1/R(1/\mathcal{R}_V)$.

If $|z|<\mathcal{R}_D$, we may therefore construct L as described above and bring the summations in (51) behind the integral. We obtain

$$\frac{D(z)-1}{z-1} = \frac{1}{2\pi\imath}\oint_L \frac{V(1/\zeta)-1}{(1-\zeta)\big(1-zR(\zeta)\big)}\,\mathrm{d}\zeta. \tag{52}$$

Substituting $z=0$ in (52), in view of $D(0)=0$, we get

$$1 = \frac{1}{2\pi\imath}\oint_L \frac{V(1/\zeta)-1}{1-\zeta}\,\mathrm{d}\zeta.$$

Using this result in (52) again, we find

$$D(z) = \frac{z}{2\pi\imath}\oint_L \frac{V(1/\zeta)-1}{1-\zeta}\cdot\frac{1-R(\zeta)}{1-zR(\zeta)}\,\mathrm{d}\zeta. \tag{53}$$

In order to further simplify the above expression, we now split the integrand into two terms, as follows:

$$\frac{V(1/\zeta)}{1-\zeta}\frac{1-R(\zeta)}{1-zR(\zeta)} - \frac{1}{1-\zeta}\frac{1-R(\zeta)}{1-zR(\zeta)}. \tag{54}$$

The latter term has no poles inside L, since L was chosen such that $\forall\zeta\in L$: $|zR(\zeta)|<1$, which implies (due to Rouché's theorem) that $1-zR(\zeta)$ has no zeros inside L, and since the simple zero of the denominator at $\zeta=1$ (if that would be inside L) is canceled by the zero of the numerator at $\zeta=1$. We conclude that the contribution of the latter term to the value of the contour integral in (53) is zero. Therefore we can rewrite (53) as

$$D(z) = \frac{z}{2\pi\imath}\oint_L \frac{V(1/\zeta)}{1-\zeta}\cdot\frac{1-R(\zeta)}{1-zR(\zeta)}\,\mathrm{d}\zeta. \tag{55}$$

Moreover, we change the integration variable in (55) to $\xi = 1/\zeta$ (which yields a factor $-1/\xi^2$ in the integrand), and we invert the integration path L into L' but still integrate in counterclockwise sense (which yields an extra factor of -1, since the inversion of L is a clockwise path). This then leads to the desired relationship between the pgfs $D(z)$ and $V(z)$. □

If the pgf $R(z)$ of the service capacities is a rational function, the general relationship (47) can be further transformed into

$$D(z) = \frac{z-1}{z} \sum_{k=0}^{m-1} \frac{V(\alpha_k(z))}{R'(1/\alpha_k(z))} \cdot \frac{\alpha_k(z)}{\alpha_k(z)-1}, \tag{56}$$

where the functions $\alpha_k(z)$ are the m zeros for ξ of $1 - zR(1/\xi)$. The relationship (56) is valid for all z for which the zeros $\alpha_k(z)$ are distinct, which is the case for all but at most $2m-1$ values of z.

Proof. If the pgf of the service capacities is a rational function, i.e., if $R(1/z)$ is given by (1), equation (47) can be rewritten as

$$D(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V(\xi)}{\xi\,(\xi-1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)}\,\mathrm{d}\xi. \tag{57}$$

We now focus on the poles of the integrand in (57) inside the contour L'. Since the service demand of each customer is at least 1 work unit, $S(0)$ must equal 0, and by (7) $V(0)$ must equal 0 as well. The zero of the factor $V(\xi)$ in the numerator of the integrand at $\xi = 0$ then ensures that the factor ξ in the denominator does not cause a pole of the integrand at $\xi = 0$. Furthermore, the factor $Q_R(\xi) - P_R(\xi)$ in the numerator ensures that the factor $(\xi - 1)$ in the denominator does not cause a pole of the integrand at $\xi = 1$. Finally, since the contour L' was chosen such that $\forall \xi \in L' : |\xi| < \mathcal{R}_V$, $V(\xi)$ has no poles inside L' either. Therefore, the only poles of the integrand in (57) inside L' are the zeros for ξ of

$$Q_R(\xi) - zP_R(\xi), \tag{58}$$

or equivalently, of

$$1 - zR(1/\xi). \tag{59}$$

Since (58) is a polynomial in ξ of degree m, (58) has exactly m zeros for ξ. We denote these zeros for a given value of z by $\alpha_k(z)$, $k = 0, 1, ..., m-1$. It can easily be seen that all these zeros lie inside L'. Indeed, the contour L' was chosen such that $\forall \xi \in L' : |zR(1/\xi)| < 1$. This implies that $|zP_R(\xi)| < |Q_R(\xi)|$, so using Rouché's theorem we can say that (58) has as many zeros inside L' as $Q_R(\xi)$. But all m zeros (counting with multiplicities) of $Q_R(\xi)$ must lie inside L', because L' was chosen such that $\forall \xi \in L' : |\xi| > 1/\mathcal{R}_R$. This means that (58) must have exactly m zeros for ξ inside L'.

To calculate the value of the contour integral in (57), we can therefore apply Cauchy's residue theorem (see e.g. [16]). Note that the zeros $\alpha_k(z)$ are not necessarily distinct. For a given value of z, let $\alpha(z)$ denote the set of zeros for ξ of (58). The pgf of $D(z)$ is then obtained as

$$D(z) = z \sum_{\xi^* \in\, \alpha(z)} \operatorname*{Res}_{\xi=\xi^*} \left[\frac{V(\xi)}{\xi\,(\xi-1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)} \right], \tag{60}$$

where the residue at a pole ξ^* with multiplicity m^* is given by

$$\frac{1}{(m^*-1)!} \lim_{\xi \to \xi^*} \frac{\mathrm{d}^{m^*-1}}{\mathrm{d}\xi^{m^*-1}} \left[(\xi - \xi^*)^{m^*} \frac{V(\xi)}{\xi\,(\xi-1)} \cdot \frac{Q_R(\xi) - P_R(\xi)}{Q_R(\xi) - zP_R(\xi)} \right]. \tag{61}$$

Since all quantities in expression (60) are known or can be calculated numerically (when z is known), this expression may be used to evaluate $D(z)$ for any z. However, due to the (m^*-1)st derivative with respect to ξ in (61), the evaluation of $D(z)$ may be difficult in practice if the zeros $\alpha_k(z)$ are not distinct.

Note that if a zero ξ of (59) has a multiplicity larger than 1, then $R'(1/\xi)/\xi^2 = 0$. Since $R'(1/\xi)/\xi^2$ is the derivative of $-R(1/\xi)$, a rational function with degree of the numerator and the denominator at most m, there are at most $2m-1$ values of ξ for which $R'(1/\xi)/\xi^2 = 0$, with at most $2m-1$ corresponding values of z (see (59)). Therefore, for all but at most $2m-1$ values of z, the zeros $\alpha_k(z)$ of (59) are distinct. For those z, or for all z if the service-capacity distribution is one of the distributions discussed in Appendix D, a substantially simpler expression for $D(z)$ is available, because we can simplify (60) to

$$D(z) = \sum_{k=0}^{m-1} V(\alpha_k(z)) \frac{\alpha_k(z)}{\alpha_k(z)-1} \frac{1-R(1/\alpha_k(z))}{R'(1/\alpha_k(z))}. \tag{62}$$

Due to (59), we moreover have that $R(1/\alpha_k(z)) = 1/z$. This allows to simplify (62) further to (56). □

Appendix C. Relationship between the pgfs $B(z)$ and $U(z)$. In this appendix, we first prove the following general relationship between the steady-state pgf $B(z)$ of the system content and the steady-state pgf $U(z)$ of the unfinished work, at the beginning of an arbitrary slot:

$$B(z) = \frac{A(z)(z-1)}{2\pi\imath(A(z)-1)} \oint_{L'} \frac{U(\xi)}{A(S(\xi))} \cdot \frac{1}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-A(z)R(1/\xi)} \cdot S(\xi) \cdot \frac{A(z)-A(S(\xi))}{z-S(\xi)}\,\mathrm{d}\xi, \tag{63}$$

where L' is a contour around the origin such that $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|A(z)R(1/\xi)| < 1$.

Proof. By combining the intermediate results (14) and (17), we get the following expression for $B(z)$:

$$B(z) = \frac{z-1}{A(z)-1} \sum_{f=0}^{\infty} \frac{D_f(A(z))}{z^{f+1}} \sum_{i=f+1}^{\infty} \mathrm{Prob}[A=i] z^i. \tag{64}$$

To derive an expression for the conditional pgf $D_f(z)$, we can now follow the same steps as explained in Appendix B for the derivation of the expression (47) for $D(z)$, except that throughout the derivation, all random variables pertaining to customer C are conditioned on $F_C = f$. It can easily be verified that all the steps of the derivation remain valid. The conditional pgfs of D_C and V_C given that $F_C = f$ are denoted by $D_f(z)$ and $V_f(z)$ respectively, while the conditional pgf of $R_C^{(k)}$ given that $F_C = f$ simply remains $R(z)^k$, since $R_C^{(k)}$ is independent of F_C. The expression we obtain for $D_f(z)$ is therefore

$$D_f(z) = \frac{z}{2\pi\imath} \oint_{L'} \frac{V_f(\xi)}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-zR(1/\xi)}\,\mathrm{d}\xi, \tag{65}$$

where L' is a contour around the origin such that $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_{V_f}$ and $|zR(1/\xi)| < 1$. The pgf $V_f(z)$ in this expression is easily obtained in terms of $U(z)$, similarly to how we obtained the expression (7) for $V(z)$ in terms of $U(z)$; the result reads

$$V_f(z) = \frac{U(z)}{A(S(z))} \cdot S(z)^{f+1}. \tag{66}$$

Substitution of (65) and (66) then leads to

$$B(z) = \frac{A(z)(z-1)}{2\pi\imath(A(z)-1)} \sum_{f=0}^{\infty} \sum_{i=f+1}^{\infty} \oint_{L'} \frac{U(\xi)}{A(S(\xi))} \cdot \frac{1}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-A(z)R(1/\xi)} \cdot \left(\frac{S(\xi)}{z}\right)^{f+1} \cdot \text{Prob}[A=i] z^i \,\mathrm{d}\xi, \tag{67}$$

where the contour L' is chosen to be the same for all terms in the sum over f, with now $\forall \xi \in L' : \mathcal{R}_R^{-1} < |\xi| < \mathcal{R}_V$ and $|A(z)R(1/\xi)| < 1$. After swapping the summations over f and i and working out the resulting finite summation over f, we obtain

$$B(z) = \frac{A(z)(z-1)}{2\pi\imath\big(A(z)-1\big)} \sum_{i=1}^{\infty} \oint_{L'} \frac{U(\xi)}{A\big(S(\xi)\big)} \cdot \frac{1}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-A(z)R(1/\xi)} \cdot S(\xi) \cdot \frac{z^i - S(\xi)^i}{z-S(\xi)} \cdot \text{Prob}[A=i] \,\mathrm{d}\xi. \tag{68}$$

To work out the infinite summation over i, it is required that $|z| < \mathcal{R}_A$ and $\forall \xi \in L' : |S(\xi)| < \mathcal{R}_A$. However, we can prove that if $|z| < \mathcal{R}_B$, then this is always the case. First, we prove that $|z| < \mathcal{R}_A$. The system content in an arbitrary slot is the sum of two independent variables: the number of customers that arrived during the previous slot and the number of customers that arrived earlier than that and are still in the system. Since the former of these two variables has pgf $A(z)$, it follows that $\mathcal{R}_B \leq \mathcal{R}_A$. Therefore, if $|z| < \mathcal{R}_B$ then $|z| < \mathcal{R}_A$. The proof that $\forall \xi \in L' : |S(\xi)| < \mathcal{R}_A$ begins with the observation from (6) that the radius of convergence of the pgf of F_C is at most $\mathcal{R}_A$. From (5) it then follows that $\mathcal{R}_V \leq \mathcal{R}_{A\circ S}$, where $\circ$ denotes function composition, i.e., $(A \circ S)(z) = A(S(z))$. Then, since L' was chosen such that $\forall \xi \in L' : |\xi| < \mathcal{R}_V$ and since $S(z)$ is an increasing function on the part of the real axis where $0 \leq z < \mathcal{R}_{A\circ S} \leq \mathcal{R}_S$, we have that $\forall \xi \in L' : |S(\xi)| \leq S(|\xi|) < S(\mathcal{R}_V) \leq S(\mathcal{R}_{A\circ S}) = \mathcal{R}_A$.

Hence, if $|z| < \mathcal{R}_B$, we can work out the infinite sum over i and as a result, we find the desired relationship between $B(z)$ and $U(z)$. □

If the pgf $R(z)$ is a rational function, the general relationship (63) can be further transformed into

$$B(z) = \frac{z-1}{A(z)} \sum_{k=0}^{m-1} \frac{U(\beta_k(z))}{R'(1/\beta_k(z))} \frac{\beta_k(z)}{\beta_k(z)-1} \frac{A(z)-A(S(\beta_k(z)))}{z-S(\beta_k(z))} \frac{S(\beta_k(z))}{A(S(\beta_k(z)))}, \tag{69}$$

where the functions $\beta_k(z)$ are defined as $\alpha_k(A(z))$, $k = 0, 1, ..., m-1$, i.e., the zeros for ξ of $1 - A(z)R(1/\xi)$. The relationship (69) is valid for all z for which these zeros $\beta_k(z)$ are distinct, which is the case for all z but an isolated set.

Proof. If $R(z)$ is rational, we can again resort to Cauchy's residue theorem to determine the contour integral in expression (63). To this end, we however first need to find out where the poles of the integrand are. In this respect, note that the factors $S(\xi)$ and $A(S(\xi))$ don't cause any poles inside L', since with (66) the expression for $B(z)$ can be rewritten as

$$B(z) = \frac{A(z)(z-1)}{2\pi\imath(A(z)-1)} \oint_{L'} \frac{V_0(\xi)}{\xi\,(\xi-1)} \cdot \frac{1-R(1/\xi)}{1-A(z)R(1/\xi)} \cdot \frac{A(z)-A(S(\xi))}{z-S(\xi)} \,\mathrm{d}\xi, \tag{70}$$

and there are no poles of $V_0(\xi)$ inside L', because by definition $\forall \xi \in L' : |\xi| < \mathcal{R}_V \leq \mathcal{R}_{V_0}$. Also note that any zero of $z - S(\xi)$ is also a zero of $A(z) - A(S(\xi))$, so this factor in the denominator causes no poles inside L' either. Since $\forall \xi \in L' : |S(\xi)| < \mathcal{R}_A$, neither does the term $A(S(\xi))$. Finally, also the factors ξ and $\xi - 1$ in the denominator do not cause any poles, since these are canceled by $V_0(\xi)$ and $1 - R(1/\xi)$ respectively. The only poles of the integrand in (66) inside L' are therefore the zeros for ξ of

$$1 - A(z)R(1/\xi), \tag{71}$$

which are, when $R(z)$ is rational, given by $\beta_k(z) \triangleq \alpha_k(A(z)), k = 0, 1, ..., m-1$. It is again easily seen that all these zeros lie inside the contour L'; they are however not necessarily distinct. For a given z, $\beta(z)$ denotes the set of distinct zeros for ξ of (20). With Cauchy's residue theorem we then get

$$B(z) = \frac{A(z)(z-1)}{A(z)-1} \sum_{\xi^* \in \beta(z)} \operatorname*{Res}_{\xi=\xi^*} \left[\frac{V_0(\xi)}{\xi\,(\xi-1)} \cdot \frac{1 - R(1/\xi)}{1 - A(z)R(1/\xi)} \cdot \frac{A(z) - A(S(\xi))}{z - S(\xi)} \right], \tag{72}$$

where the residue at a pole ξ^* with multiplicity m^* is given by

$$\frac{1}{(m^*-1)!} \lim_{\xi \to \xi^*} \frac{\mathrm{d}^{m^*-1}}{\mathrm{d}\xi^{m^*-1}} \left[(\xi - \xi^*)^{m^*} \cdot \frac{V_0(\xi)}{\xi\,(\xi-1)} \cdot \frac{1 - R(1/\xi)}{1 - A(z)R(1/\xi)} \cdot \frac{A(z) - A(S(\xi))}{z - S(\xi)} \right]. \tag{73}$$

Again, for any z this expression contains only quantities which are known or can be calculated numerically, so it can be used to calculate $B(z)$ for any z, but again this is not always practical due to the (m^*-1)st derivatives in the expression. However, it is not hard to show, similarly to how we proved that there are at most $2m-1$ values of z for which the zeros $\alpha_k(z)$ are not distinct, that the z for which the $\beta_k(z)$ are not distinct form an isolated set. This implies that for all z except an isolated set, all m zeros $\beta_k(z)$ are distinct. For these z, or again for all z if the service-capacity distribution is one of the distributions discussed in Appendix D, a significantly simpler expression for $B(z)$ is available, as (72) can be rewritten as

$$B(z) = \frac{z-1}{A(z)-1} \sum_{k=0}^{m-1} \frac{\beta_k(z)\, V_0(\beta_k(z))}{\beta_k(z) - 1} \cdot \frac{1 - R(1/\beta_k(z))}{R'(1/\beta_k(z))} \cdot \frac{A(z) - A(S(\beta_k(z)))}{z - S(\beta_k(z))}.$$

After the substitution $R(1/\beta_k(z)) = 1/A(z)$ (see (20)), this becomes

$$B(z) = \frac{z-1}{A(z)} \sum_{k=0}^{m-1} \frac{V_0(\beta_k(z))}{R'(1/\beta_k(z))} \cdot \frac{\beta_k(z)}{\beta_k(z) - 1} \cdot \frac{A(z) - A(S(\beta_k(z)))}{z - S(\beta_k(z))}. \tag{74}$$

Finally, using the definitions (66) of $V_0(z)$ and (4) of $U(z)$, we subsequently find the desired relationship between $B(z)$ and $U(z)$. □

Appendix D. **Distributions for which the zeros $\alpha_k(z)$ are distinct.** At multiple points throughout the analysis of the queueing model, the possibility that the zeros $\alpha_k(z)$ (and $\beta_k(z)$) are not distinct led to complications in the analysis, while substantial simplifications were possible under the assumption that these zeros are distinct. We showed in Appendices B and C that the sets of z values for which the zeros $\alpha_k(z)$ and $\beta_k(z)$ respectively can be non-distinct are isolated sets (and for the zeros $\alpha_k(z)$ even finite sets). Nevertheless, if the zeros $\beta_k(1)$ happen to be non-distinct then our expression (23) for the mean system content is not applicable,

and if the zeros $\beta_k(z_B)$ are non-distinct then our expression (26) for the residue C_B used in the approximation of the tail probabilities is not applicable. Luckily, as we will show in this appendix, for many service-capacity distributions, the zeros $\alpha_k(z)$ and $\beta_k(z)$ are distinct for *all* z, and can in fact be calculated explicitly. Indeed, this is the case for the following distributions:

1. Trivially, any distribution for which $m = 1$. In particular, the Bernoulli distribution, with pgf $R(z) = 1 - \mu + \mu z$, the geometric distribution, with pgf $R(z) = 1/(\mu + 1 - \mu z)$, and the shifted geometric distribution, with pgf $R(z) = z/(\mu + 1 - \mu z)$.
2. The degenerate distribution with pgf $R(z) = z^m$. For this distribution the zeros $\alpha_k(z)$ are the complex mth order roots of $1/z$, which are obviously distinct.
3. Any "composition" of two (or more) distributions for which the zeros $\alpha_k(z)$ are distinct, where the composition of two distributions with pgfs $R_1(z)$ and $R_2(z)$ and respectively m_1 and m_2 distinct zeros $\alpha_{k,1}(z)$ and $\alpha_{k,2}(z)$ is defined as the distribution with pgf $R_1(R_2(z))$. This is the distribution that results from summing N independent samples of the second distribution, where N is a random variable following the first distribution.

   This composite distribution has $m_1 m_2$ distinct zeros $\alpha_k(z)$ which are given by $\alpha_{i,2}(\alpha_{j,1}(z))$, $i = 0, 1, ..., m_2$, $j = 0, 1, ..., m_1$, since

$$R_1(R_2(1/\alpha_{i,2}(\alpha_{j,1}(z))) = R_1(1/\alpha_{j,1}(z)) = 1/z. \tag{75}$$

   It is easily seen that these zeros distinct.

Using this composition rule, many of the most commonly used distributions can be obtained: The binomial distribution ($R_1(z) = z^{m_1}$, $R_2(z) = 1 - \mu_2 + \mu_2 z$), the negative binomial distribution ($R_1(z) = z^{m_1}$, $R_2(z) = 1/(\mu_2 + 1 - \mu_2 z)$), the "bursty" distribution that has 2 possible values: 0 and m ($R_1(z) = 1 - \mu/m + \mu z/m$, $R_2(z) = z^m$), etc.

In conclusion, if the service capacity follows any of the above distributions, the zeros $\alpha_k(z)$ (and $\beta_k(z)$) are distinct and can be calculated explicitly. In this case, our results for the pgfs $D(z)$ and $B(z)$ of the steady-state customer delay and system content, expressions (13) and (21) respectively, are closed-form expressions once the poles $\mathcal{S}_R^{-1}$ and zeros $\mathcal{N}_T^-$, which are easily calculated numerically, are known.

REFERENCES

[1] J. Abate and W. Whitt, Numerical inversion of probability generating functions, *Oper. Res. Letters*, **12** (1992), 245–251.

[2] I. J. B. F. Adan and V. G. Kulkarni, Single-server queue with Markov-dependent inter-arrival and service times, *Queueing Systems*, **45** (2003), 113–134.

[3] I. J. B. F. Adan, J. S. H. van Leeuwaarden and E. M. M. Winands, On the application of Rouché's theorem in queueing theory, *Oper. Res. Letters*, **34** (2006), 355–360.

[4] S. Ayed, D. Sofiene and R. Nidhal, Joint optimisation of maintenance and production policies considering random demand and variable production rate, *International Journal Of Production Research*, **50** (2011), 6870–6885.

[5] J. W. Bosman, R. D. van der Mei and R. Nunez-Queija, A fluid model analysis of streaming media in the presence of time-varying bandwidth, *Proc. ITC 24*, 2012, 177–184.

[6] O. J. Boxma and I. A. Kurkova, The $M/G/1$ queue with two service speeds, *Advances in Applied Probability*, **33** (2001), 520–540.

[7] H. Bruneel and B. G. Kim, *Discrete-time Models for Communication Systems Including ATM*, Kluwer Academic, Boston, USA, 1993.

[8] H. Bruneel, S. Wittevrongel, D. Claeys and J. Walraevens, Discrete-time queues with variable service capacity: A basic model and its analysis, *Annals of Operations Research*, **239** (2016), 359–380.

[9] H. Bruneel, W. Rogiest, J. Walraevens and S. Wittevrongel, On queues with general service demands and constant service capacity, QEST 2014, *LNCS*, **8657** (2014), 210–225.

[10] M. Chen, X. Jin, Y. Wang, X. Q. Cheng and G. Min, Modelling priority queuing systems with varying service capacity, *Frontiers of Computer Science*, **7** (2013), 571–582.

[11] M. De Muynck, S. Wittevrongel and H. Bruneel, Analysis of discrete-time queues with general service demands and finite-support service capacities, *Annals of Operations Research*, accepted for publication, (2015), 1–26.

[12] M. De Muynck, H. Bruneel and S. Wittevrongel, Delay analysis of a queue with general service demands and phase-type service capacities, QTNA 2015, *AISC*, **383** (2015), 29–39.

[13] B. Feyaerts, S. De Vuyst, H. Bruneel and S. Wittevrongel, Performance analysis of buffers with train arrivals and correlated output interruptions, *Journal of Industrial and Management Optimization*, **11** (2015), 829–848.

[14] S. Gao and J. Wang, On a discrete-time $GI^X/Geo/1/N-G$ queue with randomized working vacations and at most J vacations, *Journal of Industrial and Management Optimization*, **11** (2015), 779–806.

[15] B. Giri, W. Yun and T. Dohi, Optimal design of unreliable production-inventory systems with variable production rate, *European Journal of Operational Research*, **162** (2005), 372–386.

[16] M. O. González, *Classical Complex Analysis*, Marcel Dekker, New York, 1992.

[17] U. C. Gupta and V. Goswami, Performance analysis of finite buffer discrete-time queue with bulk service, *Computers & Operations Research*, **29** (2002), 1331–1341.

[18] S. Halfin, Steady-state distribution for the buffer content of an $M/G/1$ queue with varying service rate, *SIAM Journal on Applied Mathematics*, **23** (1972), 356–363.

[19] X. Jin, G. Min, S. Velentzas and J. Jiang, Quality-of-service analysis of queuing systems with long-range-dependent network traffic and variable service capacity, *IEEE Transactions on Wireless Communications*, **11** (2012), 562–570.

[20] E. Kafetzakis, K. Kontovasilis and I. Stavrakakis, Effective-capacity-based stochastic delay guarantees for systems with time-varying servers, with an application to IEEE 802.11 WLANs, *Performance Evaluation*, **68** (2011), 614–628.

[21] B. Kim and J. Kim, A single server queue with Markov modulated service rates and impatient customers, *Performance Evaluation*, **83/84** (2015), 1–15.

[22] L. Kleinrock, *Queueing Systems, Volume I: Theory*, Wiley, New York, 1976.

[23] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, Society for Industrial and Applied Mathematics, 1987.

[24] S. R. Mahabhashyam and N. Gautam, On queues with Markov modulated service rates, *Queueing Systems*, **51** (2005), 89–113.

[25] I. Mitrani, *Modelling of Computer and Communication Systems*, Cambridge University Press, Cambridge, 1987.

[26] Z. Saffer and M. Telek, Analysis of BMAP vacation queue and its application to IEEE 802.16e sleep mode, *Journal of Industrial and Management Optimization*, **6** (2010), 661–690.

[27] Z. Saffer and W. Yue, $M/M/c$ multiple synchronous vacation model with gated discipline, *Journal of Industrial and Management Optimization*, **8** (2012), 939–968.

[28] T. Takine, Single-server queues with Markov-modulated arrivals and service speed, *Queueing Systems*, **49** (2005), 7–22.

[29] B. Vinck and H. Bruneel, Analyzing the discrete-time $G^{(G)}/Geo/1$ queue using complex contour integration, *Queueing Systems*, **18** (1994), 47–67.

[30] M. Vlasiou, I. J. B. F. Adan and O. J. Boxma, A two-station queue with dependent preparation and service times, *European Journal of Operational Research*, **195** (2009), 104–116.

[31] J. Walraevens, H. Bruneel, D. Claeys and S. Wittevrongel, The discrete-time queue with geometrically distributed service capacities revisited, ASMTA 2013, *LNCS*, **7984** (2013), 443–456.

[32] Y. Yao and D. Wei-Chung Miao, Sample-path analysis of general arrival queueing systems with constant amount of work for all customers, *Queueing Systems*, **76** (2014), 283–308.

Received October 2015; 1st revision March 2016; final revision June 2016.

E-mail address: MichielR.DeMuynck@UGent.be
E-mail address: Sabine.Wittevrongel@UGent.be
E-mail address: Herwig.Bruneel@UGent.be