Discrete-time queues with variable service capacity: a basic model and its analysis

Herwig Bruneel, Sabine Wittevrongel, Dieter Claeys, Joris Walraevens

Department of Telecommunications and Information Processing
Ghent University - UGent
E-mail: {hb,sw,dclaeys,jw}@telin.UGent.be

Abstract

In this paper, we present a basic discrete-time queueing model whereby the service process is decomposed in two (variable) components: the demand of each customer, expressed in a number of work units needed to provide full service of the customer, and the capacity of the server, i.e., the number of work units that the service facility is able to perform per time unit. The model is closely related to multi-server queueing models with server interruptions, in the sense that the service facility is able to deliver more than one unit of work per time unit, and that the number of work units that can be executed per time unit is not constant over time.

Although multi-server queueing models with server interruptions - to some extent - allow us to study the concept of variable capacity, these models have a major disadvantage. The models are notoriously hard to analyze and even when explicit expressions are obtained, these contain various unknown probabilities that have to be calculated numerically, which makes the expressions difficult to interpret. For the model in this paper, on the other hand, we are able to obtain explicit closed-form expressions for the main performance measures of interest. Possible applications of this type of queueing model are numerous: the variable service capacity could model variable available bandwidths in communication networks, a varying production capacity in factories, a variable number of workers in an HR-environment, varying capacity in road traffic, etc.

Key words: queueing; customer demand; service capacity; closed-form results

1 Introduction

In classical queueing models, it is generally assumed that some kind of customers require some kind of service from a given service facility, containing one or multiple servers, which are each able to provide service to one customer at a time. A large body of work exists especially on the analysis of *single-server* queues with various types of arrival processes, service times, queueing disciplines, storage capacities, etc., both in continuous-time and (to a lesser extent) discrete-time settings. The study of *multiserver* queues, however, has traditionally received much less attention, one important reason being that multiserver queues are notoriously hard to analyze. In addition, even when expressions are obtained, these contain various unknown probabilities that have to be calculated numerically, which makes the expressions difficult to interpret ([1, 13, 18]).

An important subclass of queueing models consists of queues with *server interruptions*, whereby these server interruptions may be either driven by external processes (such as breakdowns or higher-priority customers demanding service) or they may correspond to vacations deliberately taken by the servers (e.g. when the system becomes empty after a busy period). Here, again, the case of one single server subject to interruptions has been the most commonly studied setting [16, 37, 31, 39, 27], although some work has also been reported on multiserver systems with server breakdowns [18, 28, 38, 13]. The current paper[1] is related to the latter type of models in the sense that we consider a system whose service facility is able to deliver more than one unit of work per time unit (as in other multiserver models) and that the number of work units that can be executed per time unit is not constant over time (which is typical for systems with server interruptions). The key feature of the model under investigation is the decomposition of the service process in two components: *service capacity* and *customer demand*. More specifically, we consider a model in which the so-called service capacity, i.e., the number of work units that the service facility is able to perform, changes randomly from time slot to time slot. Customers demanding variable amounts of work (as specified by the service-demand distribution) enter the system and are served in First-Come-First-Served (FCFS) order, i.e., during each slot the service facility executes as many work units as possible (as specified by the momentary service capacity) to the customers present in the system, in their order of arrival. The service of a next customer is only started when the previous customer has received complete service.

The model we study is relevant in many application areas. For instance, it is closely related to the "effective bandwidth" or "effective capacity" concepts in telecommunication networks, to model the time-varying capacity of stations in wireless networks/LANs [26, 14, 25]. A wireless station can indeed be regarded as a server with varying capacity, due to rate fluctuations of the physical channel or at the MAC layer. Another potential application domain is the modeling of varying production capacity of a production system with a single product line, in order to estimate the influence of this variability on the holding times. Machines deteriorate gradually so that their production rate decelerates automatically or artificially to extend the life span of the machines. Ultimately, machines have to be repaired or replaced due to defects [2, 20, 21, 40]. In addition, in road traffic, the capacity of roads or railways is subject to factors such as accidents,

[1]This paper is an extended version of our conference paper [12].

defects, weather conditions, et cetera [34, 24].

It turns out that, although the model dealt with in this paper is related to the category of very hard multiserver-type queues with server interruptions, it allows for a completely analytic solution of the problem: no numerical calculation of unknown probabilities is required anymore. Remarkably simple explicit expressions can be derived for most quantities of interest, such as the probability generating functions (pgf's) and mean values (and, in some special cases, also the complete probability mass functions (pmf's)) of the unfinished work, the queueing delay and (with some restrictions) the buffer occupancy.

The structure of the remainder of this paper is as follows. First, we describe the model in detail in section 2. Section 3 then presents the analysis of the (steady-state) *unfinished work* in the system, resulting in an explicit expression for the pgf of this quantity. In section 4, we derive the pgf of the (steady-state) *delay* of an arbitrary customer from the pgf of the unfinished work. All these results are valid for arbitrary service-demand distributions. It turns out that the derivation of the pgf of the (steady-state) *buffer occupancy* is much harder. This is the topic of section 5. Section 6 is devoted to the special case where the service-demand distribution is geometric; in this case, we do succeed to derive an explicit expression for the pgf of the buffer occupancy. In section 7, we focus on two special arrival processes (Bernoulli and geometric), for which the whole distributions of all relevant performance measures can be obtained in explicit closed form. We discuss the results both conceptually and quantitatively in section 8. Some conclusions are drawn and directions for future work are given in section 9.

2 Mathematical model

We consider a discrete-time queueing system with infinite waiting room and a service facility (henceforth also referred to as the "server" of the system) which can deliver a variable amount of service as time goes by. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as *time slots*, *time units* or, simply, *slots*, in the sequel. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries.

The arrival process of new customers in the system is characterized by means of a sequence of i.i.d. nonnegative discrete random variables with common probability mass function (pmf) $a(n)$ and common probability generating function (pgf) $A(z)$ respectively. More specifically,

$$a(n) \triangleq \text{Prob}[\ n \text{ customer arrivals in one slot }] \quad , \quad n \geq 0 \ ,$$

$$A(z) \triangleq \sum_{n=0}^{\infty} a(n)\, z^n \ .$$

The mean number of customer arrivals per slot, in the sequel referred to as the *mean arrival rate*, is given by

$$\lambda \triangleq A'(1) \ .$$

Although the whole analysis discussed in the next sections is valid for completely arbitrary arrival distributions, we will consider further in the paper two specific choices of the arrival

process for which very explicit results can be obtained. We now also introduce these two special cases. The first case, referred to as "Bernoulli arrivals" (with mean λ), concerns the situation whereby at most one customer can arrive in the system per slot, implying $a(n)$ and $A(z)$ to be given by

$$\begin{aligned} a(0) &= 1-\lambda \quad ; \quad a(1) = \lambda \quad , \\ A(z) &= 1-\lambda+\lambda z \ . \end{aligned} \tag{1}$$

The second special case is the one of "geometric arrivals" (with mean λ), characterized by

$$\begin{aligned} a(n) &= \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^n \quad , \quad n \geq 0 \ , \\ A(z) &= \frac{1}{1+\lambda-\lambda z} \ . \end{aligned} \tag{2}$$

The service process of the customers is described in two steps. First, we characterize the *demand* that customers place upon the resources of the system, by attaching to each customer a corresponding *service requirement* or *service demand*, which indicates the number of *work units* required to give complete service to the customer at hand. The service demands of consecutive customers arriving at the system are modeled as a sequence of i.i.d. positive discrete random variables with common pmf $s(n)$ and common pgf $S(z)$ respectively. More specifically,

$$s(n) \triangleq \text{Prob[service demand equals } n \text{ work units]} \quad , \quad n \geq 1 \ ,$$

$$S(z) \triangleq \sum_{n=1}^{\infty} s(n)\, z^n \ .$$

The mean service demand of the customers is given by

$$\frac{1}{\sigma} \triangleq S'(1) \ .$$

In some of our derivations further in the paper, we will restrict ourselves to the case whereby the service demands are geometrically distributed with mean value $1/\sigma$, such that

$$\begin{aligned} s(n) &= \sigma(1-\sigma)^{n-1} \quad , \quad n \geq 1 \ , \\ S(z) &= \frac{\sigma z}{1-(1-\sigma)z} \ . \end{aligned} \tag{3}$$

Next, we describe the (variable) *resources* of the server, by attaching to each time slot a corresponding *service capacity*, which indicates the number of work units that the server is capable of delivering in this slot. We assume that service capacities are nonnegative random variables, independent from slot to slot and geometrically distributed, with common pmf $r(n)$

and common pgf $R(z)$ respectively. More specifically,

$$r(n) \triangleq \text{Prob[service capacity equals } n \text{ work units]} = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^n \quad , \quad n \geq 0 \ ,$$
$$R(z) \triangleq \sum_{n=0}^{\infty} r(n)\, z^n = \frac{1}{1+\mu-\mu z} \ . \tag{4}$$

The mean service capacity of the system (per slot) is given by

$$\mu = R'(1) \ .$$

Note that in traditional queueing models, the terms *service demand* and *service capacity* are usually not used in the sense defined here. Instead, the term *service time* is used to indicate the total time needed to serve one customer, i.e., the service time is the time a server (with an unspecified service capacity) needs to serve a customer (with an implicit service demand). Since *service time* is an ambiguous concept in our model, we do not use this term in the remainder.

The operation of the queueing system is as follows. Customers arrive in the system according to an uncorrelated arrival process, characterized by the pgf $A(z)$, as described above, and take place in the queue in their order of arrival. The amount of service required by each customer (expressed in work units) is given by their corresponding service demand, described by the pgf $S(z)$, defined above. The server serves customers from the queue one by one in FCFS order, spending no more work units in each slot than the available service capacity for that slot, which is described by the pgf $R(z)$. If the server disposes of less work units than needed to complete the service of the customer being served in a slot, the service of that customer continues in the next slot. If, on the contrary, the service capacity of the server in a slot is higher than the remaining service demand of the customer in service, then the server starts the service of the next customer in the queue (if any) or (else) becomes idle. We assume that service of a customer can start during the slot following his arrival slot at the earliest.

3 Unfinished work

3.1 System equations

We start the analysis by defining a number of important random variables. Specifically, let u_k denote the unfinished work, i.e., the total number of work units "present in" the system at the beginning of the k-th slot, and h_k the total amount of work entering the system during this slot. Furthermore, let r_k denote the service capacity during the k-th slot. Then, the following recursive system equations can be established:

$$u_{k+1} = h_k \ , \quad \text{if } u_k \leq r_k \ ,$$
$$u_{k+1} = u_k + h_k - r_k \ , \quad \text{if } u_k > r_k \ .$$

The two above cases can be summarized in one single system equation

$$u_{k+1} = h_k + (u_k - r_k)^+ \ , \tag{5}$$

by introducing the notation $(\ldots)^+$ to indicate the quantity $\max(0, \ldots)$.

In equation (5), the r_k's are a sequence of i.i.d. random variables with (known) common pgf $R(z)$, as defined in equation (4). The random variables $\{h_k\}$, on the other hand, can be obtained as

$$h_k = \sum_{i=1}^{a_k} s_{k,i} \ ,$$

where a_k indicates the number of customers entering the system during slot k (with known pgf $A(z)$), and the $s_{k,i}$'s are the service demands (with common pgf $S(z)$) of these customers (expressed in work units). It is easily seen that the h_k's are i.i.d. with pgf

$$H(z) = A(S(z)) \ . \tag{6}$$

3.2 Analysis of the unfinished work

For all k, let $U_k(z)$ denote the pgf of u_k. Then, from equation (5) we can derive

$$U_{k+1}(z) = H(z) \cdot E\left[z^{(u_k - r_k)^+}\right] \ , \tag{7}$$

with $E[\cdot]$ the expectation operator. The second factor in the right hand side of (7) can be expanded further by means of the law of total probability (using also the mutual independence of u_k and r_k):

$$E\left[z^{(u_k - r_k)^+}\right] = \sum_{n=0}^{\infty} r(n) \sum_{i=0}^{\infty} u_k(i) z^{(i-n)^+} \ ,$$

where, for all $i \geq 0$,

$$u_k(i) \triangleq \text{Prob}[u_k = i] \ .$$

Removing the $(.)^+$ operator and introducing the explicit form of the pmf $r(n)$ as defined in equation (4), we get

$$\begin{aligned}
E\left[z^{(u_k - r_k)^+}\right] &= \sum_{n=0}^{\infty} \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu}\right)^n \left(\sum_{i=0}^{n} u_k(i) + \sum_{i=n+1}^{\infty} u_k(i) z^{i-n}\right) \\
&= \sum_{i=0}^{\infty} u_k(i) \sum_{n=i}^{\infty} \frac{1}{1+\mu} (\frac{\mu}{1+\mu})^n + \sum_{i=1}^{\infty} u_k(i) z^i \sum_{n=0}^{i-1} \frac{1}{1+\mu} \left(\frac{\mu}{(1+\mu)z}\right)^n \\
&= \sum_{i=0}^{\infty} u_k(i) \left(\frac{\mu}{1+\mu}\right)^i + \frac{1}{1+\mu} \sum_{i=1}^{\infty} u_k(i) z^i \frac{(\frac{\mu}{(1+\mu)z})^i - 1}{(\frac{\mu}{(1+\mu)z}) - 1} \\
&= U_k\left(\frac{\mu}{1+\mu}\right) + \frac{z}{\mu - (1+\mu)z} \left[U_k\left(\frac{\mu}{(1+\mu)}\right) - U_k(z)\right] \ .
\end{aligned} \tag{8}$$

Combination of equations (7) and (8) then leads to

$$[\mu - (1+\mu)z]U_{k+1}(z) = H(z)\left(\mu(1-z)U_k\left(\frac{\mu}{1+\mu}\right) - zU_k(z)\right) \ . \tag{9}$$

Now, let us assume that the queueing system at hand is stable, i.e., that the stability condition [11, 36] is fulfilled. It is not difficult to see that the system is stable if and only if the mean number of work units entering the system per slot, given by $H'(1)$, is strictly less than the mean service capacity per slot, given by $R'(1)$, i.e., if and only if

$$H'(1) < R'(1) \ ,$$

or, expressed in the basic parameters of our system,

$$\frac{\lambda}{\sigma} < \mu \ . \tag{10}$$

We now let the time parameter k go to infinity. Assuming the system reaches a steady state, then both functions $U_k(\cdot)$ and $U_{k+1}(\cdot)$ converge to a common limit function $U(\cdot)$, which denotes the pgf of the unfinished work at the beginning of an arbitrary slot in steady state. As a result, equation (9) translates into a linear equation for $U(z)$, with solution

$$U(z) = \mu U\left(\frac{\mu}{1+\mu}\right) \ \frac{(z-1)H(z)}{\mu(z-1) - z[H(z)-1]} \ .$$

This expression is fully determined, except for the unknown quantity $U(\frac{\mu}{1+\mu})$, which, in turn, can be obtained by invoking the normalizing condition of the pgf $U(z)$, i.e., the condition $U(1) = 1$. It is not difficult to obtain

$$U\left(\frac{\mu}{1+\mu}\right) = 1 - \frac{\lambda}{\mu\sigma} \ . \tag{11}$$

As a final result, we then get the following closed-form expression for the pgf of the steady-state unfinished work in the system:

$$U(z) = \frac{(\mu - \frac{\lambda}{\sigma})(z-1)A(S(z))}{\mu(z-1) - z[A(S(z))-1]} \ , \tag{12}$$

where we have also used equation (6).

Various performance measures of the system, related to the unfinished work, can be derived in explicit form from the above result. For instance, the probability that the system is empty can be obtained as

$$U(0) = \left(1 - \frac{\lambda}{\mu\sigma}\right) A(0) \ . \tag{13}$$

The mean unfinished work in the system is given by

$$E[u] = U'(1) = \frac{2\lambda[(\mu+1)\sigma - \lambda] + A''(1) + \lambda\sigma^2 S''(1)}{2\sigma(\mu\sigma - \lambda)} \quad . \tag{14}$$

Higher-order moments of the unfinished-work distribution can be obtained similarly, by computing higher-order derivatives of $U(z)$.

3.3 Amount of work performed per slot

The quantities r_k indicate the *maximum* amount of work the server *can* perform during the consecutive slots, i.e., the amount of work the server *would* perform if the system had an unlimited supply of work at its disposal. Of course, during the operation of the system, the *actual* amount of work performed per slot depends on the availability of customers as time goes on. In this subsection, we therefore study the characteristics of the latter quantity.

It is clear that the amount of work performed by the server during slot k is given by

$$w_k = \min\{r_k, u_k\} \quad ,$$

where, as before, u_k denotes the total unfinished work at the beginning of slot k. It easily follows from this that the (steady-state) pgf of w_k can be computed as follows:

$$\begin{aligned} W(z) &= \sum_{n=0}^{\infty} r(n) \sum_{i=0}^{\infty} u(i) z^{\min\{n,i\}} \\ &= \sum_{n=0}^{\infty} r(n) \left(\sum_{i=0}^{n} u(i) z^i + \sum_{i=n+1}^{\infty} u(i) z^n \right) \quad , \end{aligned}$$

where

$$u(i) \triangleq \lim_{k\to\infty} u_k(i) \quad .$$

Introducing the specific form of $r(n)$ (see equation (4)), we get

$$\begin{aligned} W(z) &= \sum_{i=0}^{\infty} u(i) z^i \sum_{n=i}^{\infty} \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu}\right)^n + \sum_{i=1}^{\infty} u(i) \sum_{n=0}^{i-1} \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu}\right)^n z^n \\ &= \sum_{i=0}^{\infty} u(i) z^i \left(\frac{\mu}{1+\mu}\right)^i + \frac{1}{1+\mu-\mu z} \sum_{i=1}^{\infty} u(i) \left(1 - \left(\frac{\mu z}{1+\mu}\right)^i\right) \\ &= U\left(\frac{\mu z}{1+\mu}\right) + \frac{1}{1+\mu-\mu z} \left(1 - U\left(\frac{\mu z}{1+\mu}\right)\right) \quad . \end{aligned}$$

Amongst other performance measures, the mean amount of work executed per slot can be calculated from this and is given by

$$W'(1) = \mu \left(1 - U\left(\frac{\mu}{1+\mu}\right)\right) \ ,$$

or, in view of equation (11),

$$W'(1) = \frac{\lambda}{\sigma} \ .$$

The latter result simply confirms that the queueing system has reached a steady state, so that the average amount of work leaving the system per slot, i.e., $W'(1)$, is in balance with the average amount of work entering the system per slot, i.e., $H'(1) = \lambda/\sigma$.

4 Customer delay

We now turn to the analysis of the probability distribution of the delay (expressed in slots) customers incur in the system. More specifically, let C denote an arbitrary customer entering the system in steady state, and let J denote the slot during which C arrives. In the sequel, customer C will be referred to as the "tagged customer". We define the (discrete) delay d of customer C as the total number of (full) slots between the arrival instant of C in the system and the departure time of C from the system, i.e., d indicates the number of slots between the end of slot J and the end of the slot during which the last work unit of the service demand of C is actually being executed.

Owing to the FCFS queueing discipline used in the system, the delay d of the tagged customer C is equal to the time needed to execute the unfinished work present in the system just after slot J, but to be performed before or during the service of customer C. In the next subsections, we first compute the pgf of this amount of work. Next, from this, we derive the pgf of d.

4.1 Work to be performed before the service of the tagged customer

Let $\tilde{u}$ denote the unfinished work at the beginning of slot J, $\tilde{r}$ the available service capacity during slot J (with pgf $R(z)$ as defined in (4)), and f the number of customers entering the system during slot J but to be served before C. Then, the total amount of work to be performed before the tagged customer C, still "present in" the system just after slot J, i.e., at the moment when the delay d of customer C starts running, is given by

$$v = (\tilde{u} - \tilde{r})^+ + \sum_{i=1}^{f} \tilde{s}_i \ ,$$

where the quantities $\tilde{s}_i$ refer to the service demands of the f customers entering the system during slot J, but to be served before the tagged customer C.

It is well-known from many previous papers e.g. [3, 28, 17, 30] that the pgf of the random variable f is given by

$$F(z) \triangleq E\left[z^f\right] = \frac{A(z)-1}{\lambda(z-1)} \ . \tag{15}$$

On the other hand, the independent nature of the arrival process (from slot to slot) implies that the probability distribution of $\tilde{u}$, i.e., the unfinished work at the beginning of the *arrival slot* of the tagged customer C, is identical to the probability distribution of the unfinished work at the beginning of an *arbitrary slot* in the steady state. This implies that the pgf of $\tilde{u}$ is equal to the function $U(z)$ determined earlier (see equation (12)). For the same reason, the random variables f and $\tilde{u}$ are mutually independent. Putting all these elements together, we conclude that the pgf of v can be obtained as

$$V(z) \triangleq E[z^v] = E\left[z^{(\tilde{u}-\tilde{r})^+}\right] \cdot E\left[z^{\sum_{i=1}^{f}\tilde{s}_i}\right] = \frac{U(z)}{A(S(z))} \cdot F(S(z)) \ ,$$

where, in the last step, we have used equation (6) and the steady-state version of equation (7), i.e., equation (7) for $k \to \infty$.

Using equations (15) and (12), we can derive from the above result the following explicit expression for $V(z)$:

$$V(z) = \frac{(\mu - \frac{\lambda}{\sigma})(z-1)[A(S(z))-1]}{\lambda[S(z)-1]\left\{\mu(z-1) - z[A(S(z))-1]\right\}} \ . \tag{16}$$

4.2 Analysis of the delay

The delay d of customer C is nothing else than the number of slots required to perform the remaining service demands of the customers in front of C just after slot J (which takes v work units) together with customer C itself (which requires one full service demand $\tilde{s}$). That is, the delay d of customer C is equal to the time needed to perform $v + \tilde{s}$ work units. If we denote by $\tilde{r}_j$ the available service capacity in the j-th slot following slot J, and by q_i the total service capacity available during i consecutive slots (just after slot J), then it is not difficult to see that

$$q_i = \sum_{j=1}^{i} \tilde{r}_j \ ,$$

with corresponding pgf

$$Q_i(z) = (R(z))^i \ ,$$

with $R(z)$ as defined in equation (4).

The distribution of the delay d can then be obtained as follows. First, we express the tail distribution as

$$\text{Prob}[d > i] = \text{Prob}[q_i < v + \tilde{s}] = \sum_{n=0}^{\infty} \text{Prob}[q_i = n] \, \text{Prob}[v + \tilde{s} > n] \quad , \quad i \geq 0 \ .$$

The reasoning behind this equation is that more than i slots are required to remove $v + \tilde{s}$ units of work from the system, if and only if at most $v+\tilde{s}-1$ work units can be performed during i slots. We now z-transform the above equation and represent $\text{Prob}[q_i = n]$ as a derivative, i.e.,

$$\text{Prob}[q_i = n] = \left[\frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}Q_i(x)\right]_{x=0} = \left[\frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(R(x))^i\right]_{x=0}$$

to obtain an equation for the pgf $D(z)$ of the delay d:

$$\frac{D(z)-1}{z-1} = \sum_{i=0}^{\infty} z^i \ \text{Prob}[d > i] = \sum_{n=0}^{\infty} \frac{1}{n!} \ \text{Prob}[v + \tilde{s} > n] \left[\frac{\partial^n}{\partial x^n} \sum_{i=0}^{\infty} z^i (R(x))^i\right]_{x=0} . \qquad (17)$$

In the above equation, the partial derivative can be expressed as

$$\left[\frac{\partial^n}{\partial x^n} \sum_{i=0}^{\infty} z^i (R(x))^i\right]_{x=0} = \left[\frac{\partial^n}{\partial x^n} \frac{1}{1 - zR(x)}\right]_{x=0}$$

or, using equation (4),

$$\left[\frac{\partial^n}{\partial x^n} \sum_{i=0}^{\infty} z^i (R(x))^i\right]_{x=0} = \left[\frac{\partial^n}{\partial x^n} \frac{1+\mu-\mu x}{1-z+\mu-\mu x}\right]_{x=0} .$$

This can be further simplified into

$$\left[\frac{\partial^n}{\partial x^n} \sum_{i=0}^{\infty} z^i (R(x))^i\right]_{x=0} = \delta(n) + \frac{n!\mu^n z}{(1-z+\mu)^{n+1}} , \qquad (18)$$

where $\delta(n)$ is the well-known Kronecker delta-function, which equals 1 for $n = 0$ and 0 for all $n > 0$.

Introducing (18) in (17) then leads to

$$\frac{D(z)-1}{z-1} = 1 + \frac{z}{1-z+\mu} \sum_{n=0}^{\infty} \left(\frac{\mu}{1-z+\mu}\right)^n \text{Prob}[v + \tilde{s} > n]$$

$$= 1 + \frac{z}{1-z+\mu} \left[\frac{V(y)S(y)-1}{y-1}\right]_{y=\frac{\mu}{1-z+\mu}} .$$

Using the formula we obtained earlier for the pgf $V(\cdot)$ in equation (16), the above result translates into the following explicit expression for the pgf of the queueing delay:

$$D(z) = \frac{\mu\sigma - \lambda}{\lambda\mu\sigma} \; \frac{z(z-1)S(\frac{\mu}{1-z+\mu})\left[A(S(\frac{\mu}{1-z+\mu})) - 1\right]}{\left[S(\frac{\mu}{1-z+\mu}) - 1\right]\left[z - A(S(\frac{\mu}{1-z+\mu}))\right]} \quad . \tag{19}$$

Formula (19) expresses the pgf of the delay of an arbitrary customer in terms of known quantities only. This result is valid for arbitrary arrival distributions (with pgf $A(z)$) and arbitrary service-demand distributions (with pgf $S(z)$).

Various delay-related performance measures of the queueing system can be derived from (19) in explicit form. For instance, the mean delay can be obtained as

$$E[d] = D'(1) = 1 + \frac{\lambda\sigma + (\mu\sigma - \lambda)}{\mu\sigma(\mu\sigma - \lambda)} + \frac{\mu A''(1) + \lambda^2\sigma S''(1)}{2\lambda\mu(\mu\sigma - \lambda)} \quad . \tag{20}$$

Higher-order moments of the delay can be obtained by computing higher-order derivatives of $D(z)$ at $z = 1$.

5 Buffer occupancy

The *buffer occupancy* is defined as the number of customers present in the system. In this section we try to find an expression for the pgf and/or the mean value of the buffer occupancy from the results obtained in the previous sections.

Let us define b_k as the buffer occupancy at the beginning of the k-th slot. Then, clearly, the following relationships can be established between b_k and the unfinished work u_k, defined in section 3:

$$\begin{aligned} u_k &= 0 \quad , \quad \text{if} \quad b_k = 0 \ , \\ u_k &= \hat{s}_1 + s_2 + \ldots + s_{b_k} \quad , \quad \text{if} \quad b_k > 0 \ , \end{aligned} \tag{21}$$

where $\hat{s}_1$ indicates the remaining service demand of the customer in service at the beginning of slot k and s_2, ..., s_{b_k} denote the (full) service demands of the other $b_k - 1$ customers in the system at the beginning of slot k.

It is not straightforward to derive from the above equations a relationship between the pgf's of the unfinished work and the buffer occupancy. There are two main reasons for this. First, it is not obvious how to find the distribution (or pgf) of the random variable $\hat{s}_1$: the classical results from renewal theory [33, 10] on the distribution of the residual lifetime in a sequence of i.i.d. random variables are not applicable here, as the remaining service demand of an ongoing service does not simply decrease by one unit per slot in the system under study (because the service capacities are variable). Second, the random variables $\hat{s}_1$ and b_k, appearing in equation (21), are not necessarily independent. In the next section, however, we shall see that these obstacles do not exist if the service demands have a geometric distribution, and we shall be able to determine the pgf of the buffer occupancy completely in that case.

It should also be noted that the *mean* buffer occupancy can always be derived from the mean delay, by applying (the discrete-time version of) Little's result [15]:

$$E[b] = \lambda E[d] = \lambda + \frac{\lambda^2\sigma + \lambda(\mu\sigma - \lambda)}{\mu\sigma(\mu\sigma - \lambda)} + \frac{\mu A''(1) + \lambda^2\sigma S''(1)}{2\mu(\mu\sigma - \lambda)} \quad ,$$

where $E[d]$ was taken from equation (20).

6 Geometric service demands

In all the previous sections, we have made no specific assumptions as to the precise nature of the service-demand distribution, i.e., the pgf $S(z)$ was arbitrary. In this section, we explore the special case where the service demands are geometrically distributed with mean value $1/\sigma$, as defined in equation (3).

6.1 Unfinished work

The pgf of the unfinished work in the system at the beginning of an arbitrary slot in steady state, given in general by equation (12), cannot be much simplified in case of geometric service demands, i.e., the assumption of geometric service demands does not seem to have much impact on the form of this pgf. The result reads

$$U(z) = \frac{(\mu - \frac{\lambda}{\sigma})(z-1)A(\frac{\sigma z}{1-(1-\sigma)z})}{\mu(z-1) - z[A(\frac{\sigma z}{1-(1-\sigma)z}) - 1]} \quad . \tag{22}$$

The corresponding expression for the mean unfinished work can be obtained from this as $E[u] = U'(1)$, or, equivalently, by the substitution

$$S''(1) = \frac{2(1-\sigma)}{\sigma^2} \tag{23}$$

in equation (14). The result is given by

$$E[u] = \frac{2\lambda(1 + \mu\sigma - \lambda) + A''(1)}{2\sigma(\mu\sigma - \lambda)} \quad . \tag{24}$$

6.2 Delay

The general expression of the pgf of the customer delay, given in equation (19), simplifies considerably in case the service demands are geometrically distributed. Substitution of equation (3) in (19) yields

$$D(z) = \frac{(\mu\sigma - \lambda)z[A(\frac{\mu\sigma}{1+\mu\sigma-z}) - 1]}{\lambda[z - A(\frac{\mu\sigma}{1+\mu\sigma-z})]} \quad . \tag{25}$$

The mean delay can be found from this result as $E[d] = D'(1)$, or, from equation (20) by the substitution (23):

$$E[d] = 1 + \frac{2\lambda + A''(1)}{2\lambda(\mu\sigma - \lambda)} \quad . \tag{26}$$

6.3 Buffer occupancy

With reference to the analysis of the buffer occupancy in section 5, the assumption of geometric service demands brings about very substantial simplifications. The main reason for this lies in the memoryless nature of the geometric distribution [33, 35], which implies that the distribution of the remaining service demand ($\hat{s}_1$) is identical to the distribution of a full service demand (such as s_2, ..., s_{b_k}), i.e., the remaining service demand is also geometrically distributed with mean $1/\sigma$. Also, the distribution of $\hat{s}_1$, in this case, is not influenced by the value of b_k: although b_k may be correlated with the received amount of service of the customer in service (at the beginning of slot k), this does not determine in any way the distribution of the remaining service demand, owing to the memoryless property of the geometric distribution. As a consequence, equations (21) are equivalent to

$$u_k = \sum_{i=1}^{b_k} s_i \quad ,$$

where the s_i's are i.i.d. random variables with geometric distribution (with mean $1/\sigma$), which, in addition, are independent of the random variable b_k. It then simply follows that the (steady-state) pgf of u_k is given by

$$U(z) = B\left(\frac{\sigma z}{1 - (1-\sigma)z}\right) \quad . \tag{27}$$

The relationship

$$x = \frac{\sigma z}{1 - (1-\sigma)z}$$

can be easily inverted into

$$z = \frac{x}{\sigma + (1-\sigma)x} \quad ,$$

so that equation (27) leads to

$$B(x) = U\left(\frac{x}{\sigma + (1-\sigma)x}\right) \ .$$

Replacing x by z again and using equation (12), we then get the following explicit expression for the pgf of the buffer occupancy:

$$B(z) = \frac{(\mu\sigma - \lambda)(z-1)A(z)}{\mu\sigma(z-1) - z[A(z)-1]} \ . \tag{28}$$

The probability of an empty buffer easily follows from this as

$$B(0) = \left(1 - \frac{\lambda}{\mu\sigma}\right) A(0) \ ,$$

which is in accordance with (13) and (21). The mean buffer occupancy is easily derived as

$$E[b] = B'(1) = \lambda + \frac{2\lambda + A''(1)}{2(\mu\sigma - \lambda)} \ . \tag{29}$$

It is not difficult to see that the same result can be obtained by applying (the discrete-time version of) Little's result [15], i.e., $E[b] = \lambda E[d]$, to equation (26), which further supports our confidence in the results of our analysis.

6.4 Geometric invariance property

We conclude this section on geometric service demands with the observation that, in this particular case, the distributions of the buffer occupancy and the delay depend on the service capacities and the service demands only through the product $\mu\sigma$, as can be clearly seen from equations (25), (26), (28), (29). This is not true for the distribution of the unfinished work, though. We will refer to this remarkable property with the term "geometric invariance" in the sequel. This result implies that, for instance, doubling the mean service demands of the customers (i.e., dividing σ by 2) and at the same time doubling the mean per-slot service capacity of the system (i.e., multiplying μ by 2), does not alter the delay and the buffer occupancy of the system, but it does double the mean unfinished work in the system. Intuitively, this is very acceptable, because higher service demands appear to be compensated by the proportionally higher service capacity, and hence, the effective service times of the customers (expressed in time slots) basically remain the same. The average amount of work in the system, of course, doubles, as all work-related quantities scale up. The geometric invariance property can also be interpreted as follows: if we replace the geometric service demands (with mean $1/\sigma$ work units) with deterministic service demands equal to 1 work unit each and we slow down the server from an average service capacity of μ work units per slot to $\mu\sigma$ work units per slot, the delay and the buffer occupancy remain the same. Further, we will discover that all these conclusions are not necessarily true for non-geometric service-demand distributions.

7 Special arrival processes

The analysis in this paper did not assume any specific form for the distribution of the number of arrivals per slot. In this section, we treat the special cases of "Bernoulli arrivals" and "geometric arrivals", as introduced in section 2. It turns out that considerable simplifications of the results are possible in these two cases, especially when the service demands have a geometric distribution. We therefore only treat the latter case here.

7.1 Bernoulli arrivals

Introducing $A(z) = 1 - \lambda + \lambda z$ in equations (22) and (24) respectively, we obtain the following results related to the unfinished work:

$$U(z) = \frac{(\mu - \frac{\lambda}{\sigma})[1 - \lambda - (1 - \lambda - \sigma)z]}{\mu - [\lambda + \mu(1 - \sigma)]z}$$

and

$$E[u] = \frac{\lambda(1 + \mu\sigma - \lambda)}{\sigma(\mu\sigma - \lambda)} \quad .$$

In this case, the pgf $U(z)$ can be easily inverted; the pmf of the unfinished work is then given by

$$u(0) = \left(1 - \frac{\lambda}{\mu\sigma}\right)(1 - \lambda) \quad ;$$

$$u(i) = \left(1 - \frac{\lambda}{\mu\sigma}\right)\frac{\lambda}{\mu}\ [1 + (\mu\sigma - \lambda)]\left(1 - \sigma + \frac{\lambda}{\mu}\right)^{i-1} \quad , \quad i \geq 1 \ .$$

Similarly, the following results can be obtained with respect to the delay, from equations (25) and (26):

$$D(z) = \frac{(\mu\sigma - \lambda)z}{\mu\sigma - \lambda + 1 - z}$$

and

$$E[d] = 1 + \frac{1}{\mu\sigma - \lambda} \quad .$$

In this case, the distribution of the delay turns out to be (positive) geometric with parameter

$$\frac{1}{1 + \mu\sigma - \lambda} \quad ,$$

i.e., the pmf is given by

$$d(i) = \frac{\mu\sigma - \lambda}{1 + \mu\sigma - \lambda} \left(\frac{1}{1 + \mu\sigma - \lambda}\right)^{i-1} \quad , \quad i \geq 1 \ .$$

Finally, the pgf and mean value of the buffer occupancy, following from equations (28) and (29), are

$$B(z) = \frac{(\mu\sigma - \lambda)(1 - \lambda + \lambda z)}{\mu\sigma - \lambda z} \tag{30}$$

and

$$E[b] = \lambda + \frac{\lambda}{\mu\sigma - \lambda} \ .$$

The pmf of the buffer occupancy can be easily derived from (30) and is given by

$$b(0) = \left(1 - \frac{\lambda}{\mu\sigma}\right)(1 - \lambda) \ ;$$

$$b(i) = (1 + \mu\sigma - \lambda)\left(1 - \frac{\lambda}{\mu\sigma}\right)\left(\frac{\lambda}{\mu\sigma}\right)^{i} \quad , \quad i \geq 1 \ .$$

7.2 Geometric arrivals

Introducing

$$A(z) = \frac{1}{1 + \lambda - \lambda z}$$

in equations (22) and (24) respectively, we obtain the following results related to the unfinished work:

$$U(z) = \frac{(\mu - \frac{\lambda}{\sigma})[1 - (1 - \sigma)z]}{\mu(1 + \lambda) - [\mu(1 + \lambda - \sigma) + \lambda]z}$$

and

$$E[u] = \frac{\lambda(1 + \mu\sigma)}{\sigma(\mu\sigma - \lambda)} \ .$$

In this case, the pgf $U(z)$ can be easily inverted; the pmf of the unfinished work is given by

$$u(0) = \frac{\mu\sigma - \lambda}{\mu\sigma(1+\lambda)} \ ;$$

$$u(i) = \frac{\lambda}{\sigma} \ \frac{1+\mu\sigma}{\mu(1+\lambda-\sigma)+\lambda} \ \frac{\mu\sigma-\lambda}{\mu(1+\lambda)} \left(1 - \frac{\mu\sigma-\lambda}{\mu(1+\lambda)}\right)^i \quad , \quad i \geq 1 \ .$$

Similarly, the following results can be obtained with respect to the delay, from equations (25) and (26):

$$D(z) = \frac{(\mu\sigma - \lambda)z}{1 + \mu\sigma - (1+\lambda)z} \tag{31}$$

and

$$E[d] = \frac{1+\mu\sigma}{\mu\sigma - \lambda} \ . \tag{32}$$

Again, the distribution of the delay is (positive) geometric, in this case with parameter

$$\frac{1+\lambda}{1+\mu\sigma} \ ,$$

i.e., the pmf is given by

$$d(i) = \frac{\mu\sigma - \lambda}{1+\mu\sigma} \ \left(\frac{1+\lambda}{1+\mu\sigma}\right)^{i-1} \quad , \quad i \geq 1 \ . \tag{33}$$

Finally, the pgf and mean value of the buffer occupancy, following from equations (28) and (29), are

$$B(z) = \frac{\mu\sigma - \lambda}{\mu\sigma(1+\lambda) - \lambda(1+\mu\sigma)z}$$

and

$$E[b] = \frac{\lambda(1+\mu\sigma)}{\mu\sigma - \lambda} \ .$$

In this case, the distribution of the buffer occupancy is (nonnegative) geometric as well, more specifically with parameter

$$\frac{\lambda(1+\mu\sigma)}{\mu\sigma(1+\lambda)} \ ,$$

i.e., the pmf is given by

$$b(i) = \frac{\mu\sigma - \lambda}{\mu\sigma(1+\lambda)} \left(\frac{\lambda(1+\mu\sigma)}{\mu\sigma(1+\lambda)}\right)^i \quad , \quad i \geq 0 \ .$$

8 Discussion of results and examples

In this section, we discuss the results obtained in the previous sections, both from a qualitative perspective and by means of some examples.

The first interesting result obtained is the stability condition (10), which can be rewritten in the form

$$\lambda < \mu\sigma \ . \tag{34}$$

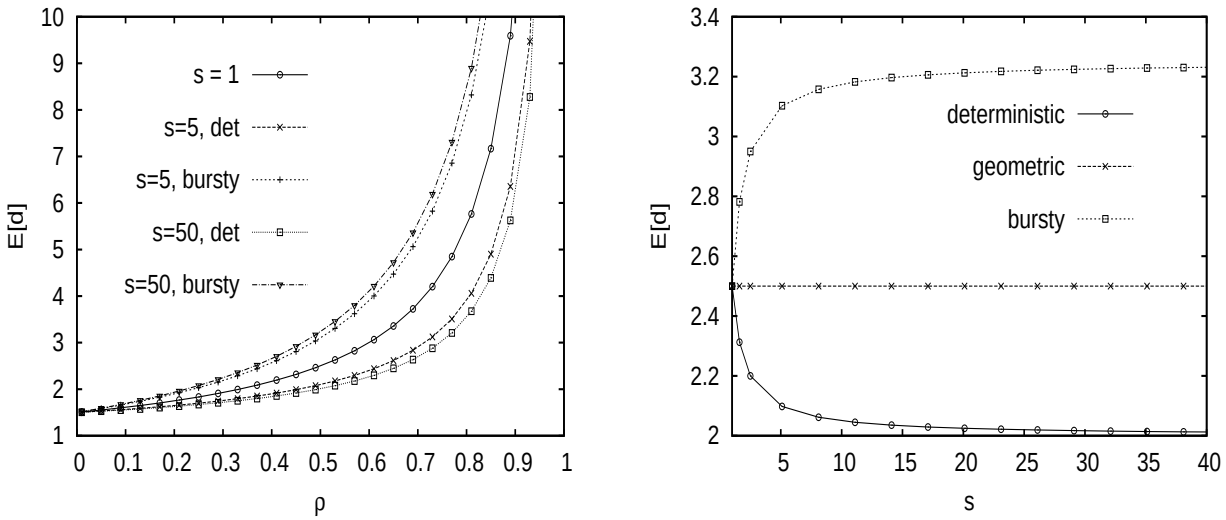
This inequality says that the supremum of the achievable throughput of the system, expressed in customers per slot, is given by $\mu\sigma$. Here λ denotes the mean number of customers entering the system per slot, μ indicates the mean number of work units that the server can deliver per slot, and σ can be viewed as the average number of customers that can be served per work unit, so that the product $\mu\sigma$ indicates the mean number of customers that can be served per slot, which gives an intuitive interpretation of the equilibrium condition.

Next, we investigate the behavior of the system when one of the parameters s ($\triangleq 1/\sigma$), μ and λ is varied. In order to make for a fair comparison, we scale one of the other parameters to keep the load ρ ($\rho = \lambda s/\mu$) constant.

First, we examine the influence of varying the mean service demand s and the mean service capacity μ while keeping their ratio constant. In Fig. 1(a), the mean delay $E[d]$ is depicted versus the load ρ, for the case of Poisson arrivals, $s/\mu = 0.5$, $s = 1, 5$ and 50 and for three possible distributions of the service demand, namely deterministic, geometric or "bursty". Here, "bursty" refers to the case where the service demand of a customer is either 1 work unit or $5{\cdot}s$ work units, i.e., where

$$S(z) = \frac{4s}{5s-1} \ z + \frac{s-1}{5s-1} \ z^{5s} \ . \tag{35}$$

When $\rho \to 0$, the mean delay tends to the constant value $1 + s/\mu$, irrespective of the absolute values of s and μ, or of the service-demand distribution. This can be understood as follows: when $\rho \to 0$, an arriving customer arrives in an empty system. He therefore has to wait 1 slot extended with, on average, an extra $1/\mu$ slots per work unit, for a total of s/μ slots. Further, the mean delay increases with increasing load ρ, as expected. For $s = 1$, the service demand is deterministically equal to 1, by definition. Because of the geometric invariance property discussed in subsection 6.4, the mean delay does not depend on the absolute values of s and μ for geometric service demands, if s/μ is kept constant. For these two reasons, the curve for $s = 1$ in Figure 1(a) is identical to the curves for other values of s and geometric service demands, as well as for $s = 1$ and the two 'other' service-demand distributions (which are, in fact, all the same for $s = 1$). When the service demands are not geometric and $s > 1$ on the other hand, we see a completely different picture; the mean delay can differ drastically when s varies, both



(a) vs. the load ρ; $s = 1, 5$ or 50 (b) vs. the mean demand s; $\rho = 0.5$

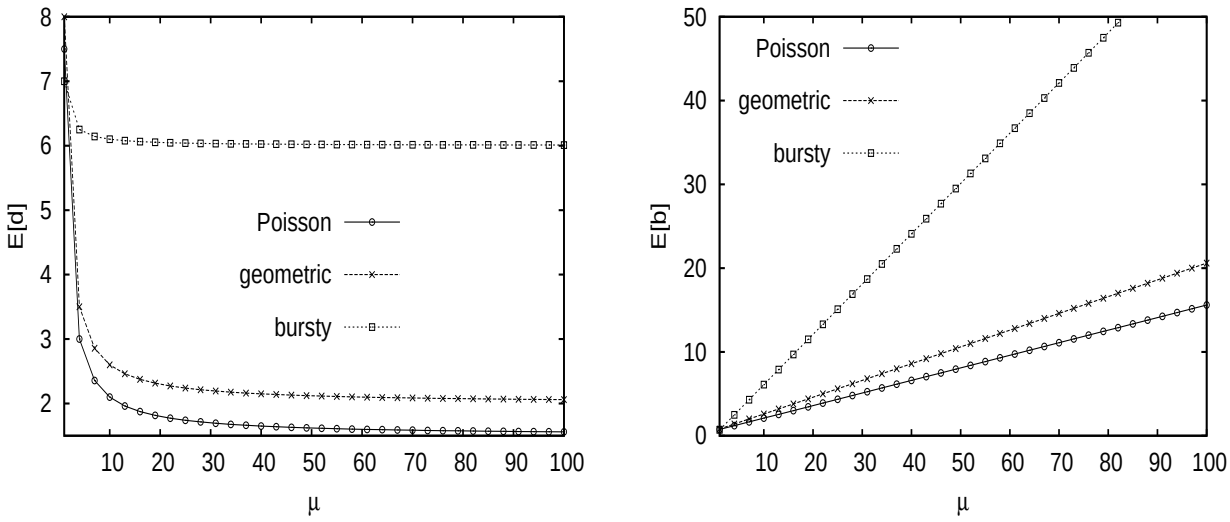
Figure 1: Mean delay $E[d]$ for Poisson arrivals and $s/\mu = 0.5$

in positive and in negative sense. For service-demand distributions with less variance than the geometric distribution (e.g., deterministic), larger service demands and larger service capacities are beneficial, while for more bursty demands, it is favorable to keep the service demands small. So, the geometric distribution seems to be a turning point. These conclusions can also be drawn from Figure 1(b). Here, we depict the mean delay versus the mean service demand s, for a constant load $\rho = 0.5$ and, otherwise, the same assumptions as in Figure 1(a). The geometric invariance property is nicely illustrated by the constant line, while it is demonstrated that less bursty (more bursty, respectively) distributions for the service demand lead to *lower* mean delay *decreasing* with s (*higher* mean delay *increasing* with s, respectively).

Second, the impact of a change in the mean arrival rate λ and the mean capacity μ is studied while maintaining their ratio constant. In Fig. 2, the mean delay and the mean buffer occupancy are plotted versus the mean service capacity μ, for deterministic service demands equal to $s = 50$, load $\rho = 0.5$ and three possible distributions for the number of per-slot arrivals, namely Poisson, geometric or "bursty". In the latter case, the number of arrivals is either 1 or 10 times the arrival rate $10 \cdot \lambda$:

$$A(z) = 0.9 + 0.1z^{10\lambda}.$$

From Fig. 2, we observe that the mean delay goes to infinity when the service capacity μ and the mean arrival rate λ go to 0, while the mean buffer occupancy tends to 0. This is logical, since an arriving customer (requiring $s = 50$ work units) needs a high number of slots to be served completely when the capacity is low (even an infinite number of slots when the service capacity goes to zero). However, the rate of arriving customers is small and, therefore, the mean buffer occupancy is small as well (and zero if λ goes to zero). So, for low μ and λ, the mean delay is almost entirely a result of one customer in the system not having enough service capacity, rather than other customers being present that have to be served before this customer.



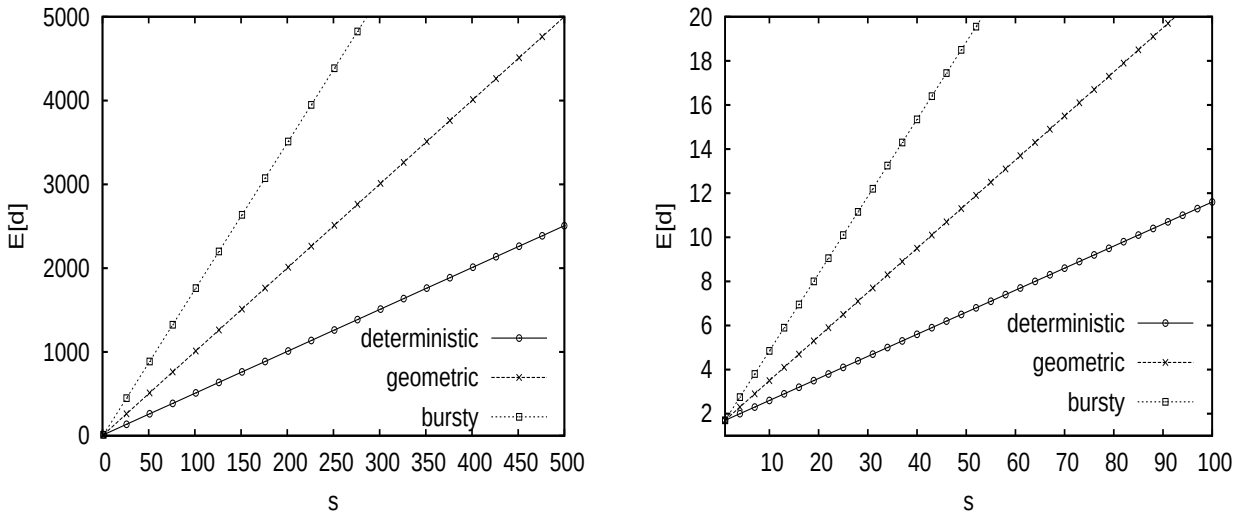
(a) Mean delay $E[d]$ (b) Mean buffer occupancy $E[b]$

Figure 2: Mean delay $E[d]$ and mean buffer occupancy $E[b]$ versus the mean service capacity μ for Poisson, geometric and bursty arrivals, deterministic service demands $s = 50$ and load $\rho = 0.5$

When the service capacity μ increases a bit, most of the customers are still waiting 'alone' in the system, so the mean delay decreases (dramatically). For even higher μ, the positive effect of increasing μ is reduced by increasing λ. For increasing λ, the mean buffer occupancy, therefore, increases, while the mean delay evolves to a constant value for μ going to ∞. We further note that a higher variance of the number of per-slot arrivals leads to higher mean delays and mean buffer occupancies.

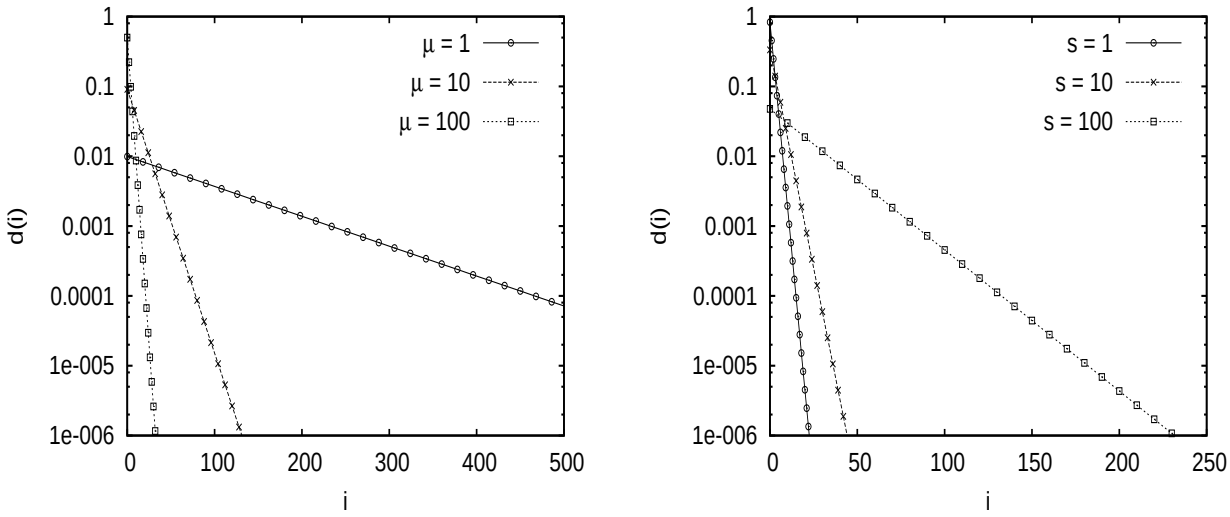
Third, we keep the mean capacity μ constant and vary the mean arrival rate λ and the mean service demand s, so as to keep the product λs constant. In Fig. 3, we depict the mean delay as a function of s, when the number of arrivals in a slot has a Poisson distribution, the service demands are deterministic, geometric or bursty (according to formula (35)), for $\rho = 0.5$ and for two values of μ, namely 0.2 (Fig. 3(a)) and 10 (Fig. 3(b)). We perceive that the mean delay increases dramatically when the mean service demand s increases, even when the mean arrival rate λ decreases inversely proportionally. This is easily explained intuitively; increasing s and decreasing λ entails less customer arrivals and a larger number of work units per customer. As a result, the work-arrival process (i.e., the sequence of work units that enter the system during consecutive slots) is burstier and thus leads to larger delays. We also notice that a higher variance of the service demands leads to a higher increment of $E[d]$ when s increases.

Finally, we turn our attention to the complete distribution of the delay. In Fig. 4, the distribution of the delay is shown in case both the service demands and the number of customer arrivals per slot have geometric distributions, for a load $\rho = 0.5$. In Fig. 4(a), we keep s constant and vary μ, while in Fig. 4(b) the converse is shown. We conclude that, for a given load ρ, the mean service capacity μ and the mean service demand s have a strong influence on the entire delay distribution; $d(i)$ decreases (much) more rapidly for low s and for high μ. This is



(a) for $\mu = 0.2$ (b) for $\mu = 10$

Figure 3: Mean delay $E[d]$ vs. the mean service demand s for λs a constant, Poisson arrivals, and $\rho = 0.5$



(a) for $s = 50$ and $\mu = 1, 10$ and 100 (b) for $\mu = 10$ and $s = 1, 10$ and 100

Figure 4: The delay distribution $d(i)$ for geometric arrivals, geometric service demands and $\rho = 0.5$

in complete accordance with the discussion on the mean delay.

9 Conclusions and future work

In this paper, we have obtained a large number of explicit expressions for the main performance measures of an elementary discrete-time queueing model with variable service capacity. These expressions allow us to study the behavior of a variable service-capacity system in conceptual terms. On the other hand, for the closely-related multi-server models with server interruptions, this is not the case: the obtained expressions always contain various unknown probabilities that have to be calculated numerically, which makes the expressions difficult to interpret. We have discovered the remarkable "geometric invariance" property for this kind of system, when the service demands are geometrically distributed. Our results demonstrate, however, that in more general conditions, changing one of the system parameters can have an undeniable impact on the system performance, even when one of the other system parameters is scaled so as to keep the system load constant.

The model examined in this paper can be generalized in various directions. To start with, *more general service-capacity distributions* can be considered than the geometric distribution. In fact, some literature exists on similar models whereby the number of work units that can be performed per slot is variable and not geometrically distributed. E.g., in [23, 22, 19], a discrete-time queueing model is examined with deterministic service times of 1 slot each and a constant number (say, m) of servers, of which a variable number is available from slot to slot. In terms of our present model, this comes down to assuming that the pgf $S(z)$ is given by $S(z) = z$ and the service-capacity distribution has *finite support*, i.e., the pgf $R(z)$ is a *polynomial* of degree m. For this model, the buffer occupancy (equivalent to the unfinished work, in this case) is examined for $m = 1$ in [23] (at arbitrary slot boundaries) and in [22] (at service-completion times), and for arbitrary $m \geq 1$ in [19] and [6], for $R(z) = 1 - \sigma + \sigma z^m$ and general $R(z)$ respectively. For the latter (most general) model, the delay analysis was performed in [28]. Although these papers consider non-geometric service-capacity distributions, they are not more general than the present study: the analysis in all these papers relies heavily on the polynomial nature of $R(z)$ and on the deterministic (single-slot) nature of the service times (two restrictions that the present model does not have). Future work may focus on a generalization to more general or even completely arbitrary distributions for the service capacity. On the other hand, the assumption that service capacities are *i.i.d.* (and, hence, *independent*) from slot to slot, may also be relaxed in future investigations. Again, some attempts in this direction have been made for the special case of single-slot service times, such as in [4, 5, 9, 32, 29] for $m = 1$, and in [7, 8] for general finite $m \geq 1$. But more general correlated models for the sequence of service capacities (from slot to slot), combined with arbitrarily distributed service times, are yet to be analyzed.

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

References

[1] 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. *Operations Research Letters*, 34:355–360, 2006.

[2] S. Ayed, S. Dellagi, and N. Rezg. Optimal integrated maintenance production strategy with variable production rate for random demand and subcontracting constraint. In *Proceedings of the 18th IFAC World Congress, 2011*, 2011.

[3] H. Bruneel. Buffers with stochastic output interruptions. *Electronics Letters*, 19:735–737, 1983.

[4] H. Bruneel. On the behavior of buffers with random server interruptions. *Performance Evaluation*, 3(3):165–175, 1983.

[5] H. Bruneel. Analysis of an infinite buffer system with random server interruptions. *Computers and Operations Research*, 11(4):373–386, 1984.

[6] H. Bruneel. A general model for the behaviour of infinite buffers with periodic service opportunities. *European Journal of Operational Research*, 16(1):98–106, 1984.

[7] H. Bruneel. A mathematical model for discrete-time buffer systems with correlated output process. *European Journal of Operational Research*, 18(1):98–110, 1984.

[8] H. Bruneel. A discrete-time queueing system with a stochastic number of servers subjected to random interruptions. *Opsearch*, 22(4):215–231, 1985.

[9] H. Bruneel. A general treatment of discrete-time buffers with one randomly interrupted output line. *European Journal of Operational Research*, 27(1):67–81, 1986.

[10] H. Bruneel. Performance of discrete-time queueing systems. *Computers and Operations Research*, 20(3):303–320, 1993.

[11] H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.

[12] H. Bruneel, J. Walraevens, D. Claeys, and S. Wittevrongel. Analysis of a discrete-time queue with geometric service capacities. In *Proceedings of the 19th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA'12)*, pages 121–135, Grenoble, 2012.

[13] S.R. Chakravarthy. Analysis of a multi-server queue with Markovian arrivals and synchronous phase type vacations. *Asia-Pacific Journal of Operational Research*, 26(1):85–113, 2009.

[14] C.S. Chang and J.A. Thomas. Effective bandwidth in high-speed digital networks. *IEEE Journal on Selected Areas in Communications*, 13:1091–1100, 1995.

[15] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.

[16] D. Fiems, B. Steyaert, and H. Bruneel. Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 35:277–298, 2004.

[17] P. Gao, S. Wittevrongel, and H. Bruneel. Discrete-time multiserver queues with geometric service times. *Computers and Operations Research*, 31:81–99, 2004.

[18] P. Gao, S. Wittevrongel, K. Laevens, D. De Vleeschauwer, and H. Bruneel. Distributional Little's law for queues with heterogeneous server interruptions. *Electronics Letters*, 46:763–U51, 2010.

[19] N.D. Georganas. Buffer behavior with poisson arrivals and bulk geometric service. *IEEE Transactions on Communications*, 24:938–940, 1976.

[20] B.C. Giri, W.Y. Yun, and T. Dohi. Optimal design of unreliable productioninventory systems with variable production rate. *European Journal of Operational Research*, 162(2):372386, 2005.

[21] C.H. Glock. Batch sizing with controllable production rates. *International Journal of Production Research*, 48:5925–5942, 2010.

[22] T.S. Heines. Buffer behavior in computer communication systems. *IEEE Transactions on Computers*, 28:573–576, 1979.

[23] J. Hsu. Buffer behavior with poisson arrival and geometric output process. *IEEE Transactions on Communications*, 22:1940–1941, 1974.

[24] Z. Hwang, J. Kim, and S. Rhee. Development of a new highway capacity estimation method. In *Proceedings of the Eastern Asia Society for Transportation Studies, 2005*, pages 984–995, 2005.

[25] X. Jin, G. Min, and S.R. Velentzas. An analytical queuing model for long range dependent arrivals and variable service capacity. In *Proceedings of IEEE International Conference on Communications (ICC 2008)*, pages 230–234, Beijing, May 2008.

[26] 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*, 68:614–628, 2011.

[27] F. Kamoun. Performance evaluation of a queuing system with correlated packet-trains and server interruption. *Telecommunication Systems*, 41(4):267–277, 2009.

[28] K. Laevens and H. Bruneel. Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers. *European Journal of Operations Research*, 85:161–177, 1995.

[29] D.-S. Lee. Analysis of a single server queue with semi-Markovian service interruption. *Queueing Systems*, 27:153–178, 1997.

[30] T. Maertens, J. Walraevens, and H. Bruneel. A modified HOL priority scheduling discipline: performance analysis. *European Journal of Operations Research*, 180:1168–1185, 2007.

[31] C. Malchin and H. Daduna. Discrete time queueing networks with product form steady state. Availability and performance analysis in an integrated model. *Queueing Systems*, 65:385–421, 2010.

[32] M. Mehmet Ali, X. Zhang, and J.F. Hayes. A performance analysis of a discrete-time queueing system with server interruption for modeling wireless atm multiplexer. *Performance Evaluation*, 51:1–31, 2003.

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

[34] B. Morris, S. Notley, K. Boddington, and T. Rees. External factors affecting motorway capacity. *Procedia Social and Behavioral Sciences*, 16:69–75, 2011.

[35] A. Papoulis and S.U. Pillai. *Probability, random variables, and stochastic processes, 4th Edition*. Mc Graw-Hill, New York, USA, 2002.

[36] H. Takagi. *Queueing Analysis, A Foundation of Performance Evaluation, Volume 3: Discrete-time systems*. North-Holland, Amsterdam, The Netherlands, 1993.

[37] T. Takine and B. Sengupta. A single server queue with service interruptions. *Queueing Systems*, 26:285–300, 1997.

[38] X.L. Yang and A.S. Alfa. A class of multi-server queueing system with server failures. *Computers & Industrial Engineering*, 56(1):33–43, 2009.

[39] M.M. Yu, Y.H. Tang, Y.H. Fu, and L.M. Pan. An M/E(k)/1 queueing system with no damage service interruptions. *Mathematical and Computer Modelling*, 54(5-6):1262–1272, 2011.

[40] H. Zied, D. Sofiene, and R. Nidhal. An optimal production/maintenance planning under stochastic random demand, service level and failure rate. In *Proceedings of the IEEE International Conference on Automation Science and Engineering, 2009 (CASE 2009)*, pages 292–297, 2009.