The discrete-time queue with geometrically distributed service capacities revisited

Joris Walraevens[⋆], Herwig Bruneel, Dieter Claeys, and Sabine Wittevrongel

Department of Telecommunications and Information Processing
Ghent University - UGent
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.
Phone: +32-9-2648902, Fax: +32-9-2644295
E-mail: {jw,hb,dclaeys,sw}@telin.UGent.be

Abstract. We analyze a discrete-time queue with variable service capacity, such that the total amount of work that can be performed during each time slot is a stochastic variable that is geometrically distributed. We study the buffer occupancy by constructing an analogous model with fixed service capacity. In contrast with classical discrete-time queueing models, however, the service times in the fixed-capacity model can take the value zero with positive probability (service times are non-negative). We study the late arrival models with immediate and delayed access, the first model being the most natural model for a system with fixed capacity and non-negative service times and the second model the more practically relevant model for the variable-capacity model.

1 Introduction

In this manuscript, we are interested in the study of a discrete-time queue with variable service capacity. In classical discrete-time queueing systems, the service capacity is fixed, equal to 1 in single-server models (e.g. [1, 2]), and larger than 1 in multiserver models (e.g. [3, 4]). Furthermore, service capacity is expressed in number of customers that can be served simultaneously. Therefore, the *service time*, the number of time slots a customer is effectively being served, is equal to the *service requirement*, the total amount of work units his service requires. In the current model, however, a variable number of *work* units can be executed by the server in one time unit, and this in a purely FCFS (First Come First Served) manner, resulting in a service time that can be different from the service requirement of a customer. We assume the service capacity to be geometrically distributed, independent from slot to slot.

This paper is a follow-up paper to [5], where the unfinished work and the customer delay were analyzed for the same model. The buffer occupancy, however, was only obtained in case of geometric service requirements, as difficulties in analyzing the buffer occupancy were observed when service requirements have a general distribution. We here solve this latter problem, by looking at

[⋆] Corresponding author

the problem from a different angle. We first construct an analogy with a fixed service-capacity model with non-negative service times. We further show that, for the latter model, the same formulas can be found for the probability generating functions (pgfs) of buffer occupancy and customer delay as in a model with strictly positive service times, a model that is commonly adopted (see e.g. [2]). Therefore, a remarkable side result of our analysis is that the adoption of non-zero service times seems to be an unnecessary assumption in discrete-time queueing analyses (although analyses are usually heavily based on it).

The model we study is relevant in many application areas. We illustrate this through discussion of two important example domains. First, 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 [6–8]. 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. A second potential application domain is the modeling of varying production capacity of a production system with a single product line [9–11], in order to estimate the influence of this variability on the holding times.

The remainder of this paper is composed as follows. In section 2, the model is introduced in detail. The translation to a fixed service-capacity model is tackled in section 3. Next, we calculate the pgf of the buffer occupancy in section 4, and analyze moments and asymptotic probabilities in section 5. In section 6, we shortly comment on the customer delay, which although retrieved in full in [5], can also be found more elegantly with our approach. Some conclusions are drawn in section 7.

2 Model

We review the model of [5] in this section. We consider a discrete-time queueing system with infinite waiting room and a server which can deliver a variable amount of service as time goes by, called the service capacity. We assume that all events (arrivals and departures) occur at the end of slots. The arrival process of new customers in the system is characterized by means of a sequence of i.i.d. (independent and identically distributed) non-negative discrete random variables with common probability mass function (pmf) $a(n)$ and common probability generating function (pgf) $A(z)$ respectively. More specifically,

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

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

The mean number of customer arrivals per slot is given by

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

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 requirements 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,

$$\begin{aligned} s(n) &\triangleq \mathrm{Pr}[\text{service demand equals } n \text{ work units}], \quad n \geq 1, \\ S(z) &\triangleq \sum_{n=1}^{\infty} s(n) z^n. \end{aligned}$$

The mean service demand of the customers is given by

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

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,

$$\begin{aligned} r(n) &\triangleq \mathrm{Pr}[\text{service capacity equals } n \text{ work units}] \\ &= \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu} \right)^n, \quad n \geq 0, && (1) \\ R(z) &\triangleq \sum_{n=0}^{\infty} r(n) z^n \\ &= \frac{1}{1+\mu-\mu z}. && (2) \end{aligned}$$

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

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

3 Translation to a fixed-capacity model

Up to now, we have not yet defined the concept *service time of a customer*, the number of slots a customer resides in the server. We have not even elaborated on the concept 'the server'. We note that this can be done in different ways. For instance, we could assume that all customers that are (partly) served during a slot are residing in the server during that slot; this slot would then count as a slot of the service times of all these customers and these service times would hence be overlapping. We opt here for a different approach. We assume that customers are served one by one and that the service time of a customer is the number of slots it resides in the server. Service times of consecutively served customers are therefore non-overlapping, as in a regular single-server model. Note that we assumed that all events occur at the end of the slots. Thus, in order to be able to model the fact that more than 1 customer can end service in the same slot, we have to assume that service times can be equal to zero slots. We clarify this with an example. Say that three customers are in the system in a given slot and that the remaining service requirement of the first customer equals 3 work units, while the service requirements of the other two equal 5 and 4 work units respectively. The slot then counts as a slot of the service time of the first customer only, regardless of the service capacity in that slot. If the service capacity in the slot is higher than 7, both the first and second customer leave the system at the end of the same slot, and the service time of the second customer is therefore assumed to be zero.

As said, service times of different customers are non-overlapping. Due to the i.i.d. service capacities and the memoryless nature of the geometric distribution, service times of consecutively served customers are also independent. Indeed, when a customer departs at the end of slot k (say), the remaining service capacity in slot k is spent on serving the following customer (and possibly even other customers). The remaining part of a random variable with a geometric distribution is independent of the elapsed part, and has the same distribution as the random variable itself. Furthermore, service capacities in later slots are independent of the service capacity in slot k, and therefore the actual service time of the second customer is independent of that of the first customer.

We now calculate the pgf of the service time v of a random customer, i.e., the required number of slots to serve a customer completely. Assume a random customer. Since the service capacity is geometrically distributed with parameter $\mu/(1+\mu)$ (as in (1)), the service time of this customer can be written as

$$v = \sum_{i=1}^{s} v_i, \tag{3}$$

with s the service requirement of the tagged customer and v_i the number of slots required to process the i-th work unit of this customer. Due to the geometric distribution of the service capacity (and its memoryless property), the service capacity in a slot has expired with probability $1/(1+\mu)$, or, capacity of at least one work unit remains with probability $\mu/(1+\mu)$, irrespective of the elapsed part of the service capacity. Therefore, the v_i's are independent and are all geometrically distributed with parameter $1/(1+\mu)$:

$$\begin{aligned}
\Pr[v_i = n] &= \frac{\mu}{1+\mu}\left(\frac{1}{1+\mu}\right)^n, \quad n \geq 0, \\
V_i(z) &\triangleq \mathrm{E}[z^{v_i}] \\
&= \frac{\mu}{1+\mu-z}.
\end{aligned}$$

Since the v_i's are independent of s as well, (3) leads to

$$V(z) = S\left(\frac{\mu}{1+\mu-z}\right). \tag{4}$$

We may conclude that our queueing system is equivalent to a discrete-time single-server system with a fixed service capacity of one work unit per slot, with the (uncommon) characteristic that service times can be zero, as the pgf of the service times is given by (4); specifically, the probability that the service time is zero equals

$$V(0) = S\left(\frac{\mu}{1+\mu}\right).$$

For further use, we also define the mean service time:

$$\nu = V'(1) = \frac{1}{\sigma\mu}. \tag{5}$$

The analysis in the remainder makes extensive use of this analogy, in contrast with the analysis in [5]. In [5], first the unfinished work at random slot boundaries was analyzed, and from this unfinished work, buffer occupancy and customer delay were investigated. This is a primarily complex-analytic analysis and the analysis of the buffer occupancy is restricted to *geometric* service requirements (in addition to geometric service capacities). Here, we resort to a more intuitive, stochastic and direct analysis of the buffer occupancy and customer delay, for a *general* service-requirement distribution.

4 Buffer Occupancy

From the previous section, we conclude that we can resort to the analysis of a regular discrete-time single-server model with non-negative service times, the latter variable's distribution being characterized by a general pgf $V(z)$. Substitution of (4) in the resulting expressions will then finally deliver results for the variable service-capacity model.

The fact that service times might be zero is a complication. For instance in [2], the positiveness of the service times is explicitly used in the analysis, which makes those results invalid for our model. However, we will retrieve the same results for non-negative service times, through a different analysis.

We first analyze the buffer occupancy in the late arrival model with immediate access [12, 13], which is from a theoretical point of view the most natural model for a fixed service-capacity queueing system with non-negative service times (we will comment on this later). In a second model, we then adopt the model with delayed access, as in the variable service-capacity model of [5], which is a more practically-oriented model. We will use results of the first model in the analysis of the latter.

4.1 Immediate Access

We assume the following order of events at the end of each slot: first (i) customers arrive, then (ii) customers leave and finally (iii) we observe the buffer occupancy. Customers that have just arrived might depart directly, thus leading to zero customer delays. We first analyze the buffer occupancy as seen by a departing customer. We then relate this to the buffer occupancy as seen by an arriving customer and conclude with the analysis of the buffer occupancy at a random slot mark.

Denote the buffer occupancy seen by the k-th departing customer as $\tilde{b}_k$, $k \geq 1$. Note that, if customer $k+1$ departs at the same time as customer k, we count customer $k+1$ as part of $\tilde{b}_k$. The service time of customer k equals v_k and the number of customers arriving in the same batch as customer k is denoted by $\tilde{a}_k$. The number of customer arrivals during the i-th slot of the service time of customer k is given by $\tilde{a}_{k,i}$, $k \geq 1$. We then obtain following system equation:

$$\tilde{b}_{k+1} = \begin{cases} \tilde{b}_k - 1 + \sum_{i=1}^{v_{k+1}} \tilde{a}_{k+1,i} & \text{if } \tilde{b}_k > 0 \\ \tilde{a}_{k+1} - 1 + \sum_{i=1}^{v_{k+1}} \tilde{a}_{k+1,i} & \text{if } \tilde{b}_k = 0 \end{cases}, \tag{6}$$

$k \geq 1$. This is identical to equation (1.41) in [13], and thus the pgf of the steady-state buffer occupancy seen by a random departure equals (1.43) in the same book:

$$\tilde{B}(z) = \frac{1-\lambda\nu}{\lambda} \cdot \frac{(A(z)-1)V(A(z))}{z - V(A(z))}. \tag{7}$$

In a second step, we relate the pgfs of the buffer occupancies as seen by departing and arriving customers. As for the departures, we will assume that, if more than one customer arrives at the same time, there is a precise order in the arrival sequence. So, if the k-th and $(k+1)$-st arriving customers arrive at the same epoch, the $(k+1)$-st customer sees the k-th customer as part of the buffer occupancy upon arrival, while the opposite is not true. Since we therefore look at the buffer occupancy as left by departing customers and as seen by arriving customers (rather than at arrival and departure epochs), we can regard the system as a single-arrival, single-departure system (see [14] for a similar observation). For this type of systems, it holds that buffer occupancies as seen by departing and arriving customers are identically distributed (see [15]), and thus

$$\hat{B}(z) = \tilde{B}(z), \tag{8}$$

with $\hat{B}(z)$ the pgf of the buffer occupancy as seen by an arriving customer.

Finally, we connect the pgfs of the buffer occupancies as seen by an arriving customer and at random slot boundaries. We can write

$$\hat{b} = b + f, \tag{9}$$

with $\hat{b}$ the buffer occupancy as seen by a random arriving customer, b the buffer occupancy at the preceding slot boundary and f the number of customers arriving at the same epoch as the randomly selected customer and buffered 'before' him. It is clear that b and f are independent random variables. Since the pgf $F(z)$ of f is well-known (see e.g. [1]),

$$F(z) = \frac{A(z) - 1}{\lambda(z-1)}, \tag{10}$$

we find

$$\begin{aligned} B(z) &= \frac{\hat{B}(z)}{F(z)} \\ &= \frac{\lambda(z-1)\hat{B}(z)}{A(z)-1} \\ &= \frac{\lambda(z-1)\tilde{B}(z)}{A(z)-1} \\ &= (1-\lambda\nu)\,\frac{(z-1)V(A(z))}{z - V(A(z))}, \end{aligned} \tag{11}$$

where we consecutively used (9), (10), (8) and (7). Substitution of (4) in (11) yields the pgf of the steady-state buffer occupancy for the varying service-capacity model with immediate access:

$$B(z) = (1-\lambda\nu)\,\frac{(z-1)S\left(\dfrac{\mu}{1+\mu-A(z)}\right)}{z - S\left(\dfrac{\mu}{1+\mu-A(z)}\right)}.$$

We remark that the final expression (11) is the same expression as the one of the pgf of the buffer occupancy in a discrete-time Geo$^X/G/1$ queue with pgf $A(z)$ for the number of arrivals at a slot boundary, $V(z)$ the pgf of *strictly positive* service times ($V(0) = 0$) and *delayed access*, see e.g. [2]. Therefore, in our opinion, the latter case can be regarded as special case of the system analyzed in this subsection. In fact, we regard the term 'late arrival with delayed access' wrongly chosen in fixed service-capacity systems, as the late arriving customer can in fact enter the service unit immediately to start service, only his service time is at least one. Therefore, it would be better to use the terms 'systems with positive service times' or (the more general) 'systems with non-negative service times' instead.

4.2 Delayed Access

Here, we return to the model of [5]. In the model with delayed access, customers cannot leave the system immediately upon arrival. Therefore, we change the order of arrivals and departures (departures now occur just before new customers arrive, and then the system is observed), so that new arrivals have to wait at least one slot for the next service opportunity. In this model, the delay of customers is at least one slot, which is a natural assumption in practice (in contrast with the model with immediate access). This does not mean that service times of customers are at least one slot as well. Multiple customers can still leave at a particular slot boundary and the service times of all but the first are then equal to zero. For this reason, the delayed-access model is not as natural as the immediate-access model, from a single-sever fixed-capacity point of view. It also complicates the analysis slightly, but the system can still be regarded as a discrete-time fixed service-capacity system, only it is a model with delayed access.

We take the same approach as in previous subsection; we analyze the buffer occupancy respectively as seen by departing customers, as seen by arriving customers and at slot boundaries. In fact, the same system equations (6) for and pgf (7) of the buffer occupancy as seen by departing customers, and the same relation (8) between the distributions of the buffer occupancy as seen by departing and by arriving customers hold. The only difference is the relation (9) between the buffer occupancy as seen by arriving customers and the buffer occupancy at the preceding slot boundary. For the current model, we find

$$\hat{b} = b - c + f,$$

with $\hat{b}$ the buffer occupancy as seen by a random arrival, b the buffer occupancy at the previous slot boundary, f the number of arrivals at the arrival epoch of the customer and admitted before the tagged arrival, and c the total number of departures directly preceding these arrivals. The term $-c$ is thus the difference with the previous model (see (9)). Note that c and b are not independent which complicates analysis. We have

$$\hat{B}(z) = \mathrm{E}\left[z^{b-c}\right] F(z). \tag{12}$$

We calculate $\mathrm{E}\left[z^{b-c}\right]$. We can write

$$b^* = b - c + a, \tag{13}$$

with b the buffer occupancy at a random slot boundary, and c, a and b^* the number of departures, arrivals and customers in the system at the following slot boundary. Since b^* is also distributed as the buffer occupancy at a random slot boundary, we may write, from (13),

$$B(z) = \mathrm{E}\left[z^{b-c}\right] A(z). \tag{14}$$

Using (14), (12), (8), (7) and (10), we obtain the pgf $B(z)$ of the buffer occupancy at random slot boundaries as

$$B(z) = \ (1 - \lambda\nu)\,\frac{A(z)(z-1)V(A(z))}{z - V(A(z))}. \tag{15}$$

Substitution of (4) in (15) yields the pgf of the steady-state buffer occupancy for the varying service-capacity model with delayed access:

$$B(z) = (1-\lambda\nu)\frac{A(z)(z-1)S\left(\dfrac{\mu}{1+\mu-A(z)}\right)}{z-S\left(\dfrac{\mu}{1+\mu-A(z)}\right)}. \tag{16}$$

For the special case of geometric service requirements, we obtain the same result as in [5].

5 Performance Measures and Discussion

In this section, we calculate and discuss some performance measures. We concentrate on the model with delayed access. In the discussion, we will focus on the influence of the mean service capacity μ, while keeping the mean service time ν constant (i.e., by scaling the mean service demand $1/\sigma$ accordingly, see (5)). If feasible, we also assume the normalized moments (such as the coefficient of variation and skewness) of mean number of per-slot arrivals and service demand to be constant.

5.1 Moments

First, the mean buffer occupancy is found by taking the first derivative of (16) in $z=1$:

$$\mathrm{E}[b] = B'(1) = \lambda + \frac{\lambda\nu\left[1+\lambda\left(C_A^2+1\right)+\lambda\nu\left(C_S^2-1\right)\right]}{2(1-\lambda\nu)} + \frac{\lambda^2\nu}{2(1-\lambda\nu)\mu}, \tag{17}$$

where we introduced the coefficients of variation C_A and C_S of the number of per-slot arrivals and the service demands respectively. From this formula, we can conclude that if the service capacity μ goes to infinity (while scaling the mean and variance of the service demand in order to keep the mean service time ν per customer and the coefficient of variation C_S of the service demand constant), the mean buffer occupancy tends to a constant. This constant depends on the mean service time, the arrival rate and the coefficients of variations of the numbers of arrivals per slot and the service demands. If the mean service capacity goes to 0, the mean buffer occupancy tends to infinity, even though the service demand is scaled accordingly, and this according to a $1/\mu$-rule with a prefactor depending on both the arrival rate and the mean service time.

Higher moments of the buffer occupancy can be calculated by taking higher derivatives of $B(z)$ in 1. The expression for the variance, for instance, contains a constant term, a term in $1/\mu$ and a term in $1/\mu^2$. The constant term depends on the mean value, the coefficient of variation and the skewness of both the number of arrivals per slot and the service demand. The term in $1/\mu$ only depends on the mean values and coefficients of variations, while the term in $1/\mu^2$ only depends on the mean values. For reference, we give the latter as it is the dominant term for the mean service capacity μ going to 0:

$$\mathrm{Var}[b] \sim \frac{\lambda^3\nu(8-5\lambda\nu)}{12(1-\lambda\nu)^2\mu^2},$$

for $\mu \to 0$.

5.2 Asymptotic Probabilities

Asymptotics of probabilities $\Pr[b=n]$, $n \to \infty$ can be calculated by investigation of the dominant singularity of the pgf $B(z)$ (i.e. the singularity of lowest norm) and the behavior of $B(z)$ in the neighbourhood of this dominant singularity [16, 17]. This dominant zero is real and bigger than or equal to 1. The position of the dominant singularity is dependent on complete distributions, i.e., of $A(z)$ and $S(z)$. We will therefore investigate a special case instead of analyzing it in full generality. We consider the case that the functions $A(z)$ and $S(z)$ are meromorphic.

The dominant singularity z_0 of $B(z)$ is a single zero of the denominator, i.e.,

$$z_0 = S\left(\frac{\mu}{1+\mu-A(z_0)}\right), \tag{18}$$

and the corresponding probabilities decay geometrically

$$\Pr[b=n] \sim (1-\lambda\nu)\,\frac{A(z_0)(z_0-1)}{S'\left(\dfrac{\mu}{1+\mu-A(z_0)}\right)\dfrac{\mu A'(z_0)}{(1+\mu-A(z_0))^2}-1}z_0^{-n}.$$

We now investigate the decay rate z_0 in terms of μ. We vary μ, while scaling σ in order to keep a constant mean service time ν. Since z_0 has to be calculated numerically, in general, cf. (18), we look at the following illustrative example. Assume Bernoulli arrivals, i.e., in each slot one customer arrives with probability λ and no customers arrive with complementary probability $1-\lambda$. The distribution of the service demand is chosen such that we can model deterministic as well as bursty demands, namely we assume that the service demand is either 1 slot or s slots, respectively with probability p and $1-p$. The arrival rate λ is equal to 0.8, while the mean service time ν is equal to 0.5. Then μ varies linearly with s according to following relation:

$$\mu = \frac{p+(1-p)s}{\nu}$$

In table 1, we list the values of z_0 for several combinations of p and s.

For $s=1$, deterministic service demands of 1 slots each are required. In this case the probabilities decay with factor 2.5. This can be taken as reference case, as it is also the decay rate of a more general geometric distribution of the service demands. From [5], it is known that z_0 is indeed independent of the actual

s	1	2	3	4	5
$p = 0$	2.5	3.2087	3.6117	3.8705	4.0505
$p = 0.3$	2.5	2.9542	3.1771	3.3093	3.3970
$p = 0.6$	2.5	2.7155	2.7621	2.7655	2.7580
$p = 0.9$	2.5	2.5254	2.4387	2.3203	2.2103

Table 1. Asymptotic decay rate z_0 of probabilities $\Pr[b = n]$

value of μ, as long as $\nu = 1/\mu\sigma$ is fixed, in case of geometric service demands. Therefore, for increasing s, we can conclude on the impact of 'burstiness' (or lack thereof) on the decay rate of the mass function of the buffer occupancy, by comparing with the constant rate 2.5 of this geometric distribution. We see that for small p, the decay rate increases with s meaning that the probabilities decay faster for higher s. For $p = 0$, the service demand is not bursty at all, as it is deterministically equal to s. Higher service demands and service requirements are in this case advantageous. For higher p, one can see the impact of burstiness. For $p = 0.6$, the decay rate is almost constant, as is the case for the geometric distribution. For even higher p (and high s) the decay rate decreases for increasing s, which means that the probability tail decreases less fast. Therefore, lower service demands and service requirements are advantageous in this case. Another way to look at it is the following: if we assume that geometric capacity is perfectly fitted to geometric demands, we can conclude that geometric capacity is not fitted well to either smoother (too much capacity at times) or burstier (too less capacity at times) service demands.

6 Customer Delay

Let us now briefly turn to the customer delay. We assume a FCFS (First Come First Served) scheduling discipline. Customer delay is defined as the number of slots a customer sojourns in the system. The pgf of the customer delay was calculated in [5] for the model with delayed access, by means of an elaborate algebraic analysis (first the unfinished work was analyzed and customer delay was then related to the unfinished work). We here present a stochastic analysis by making use of the analogy with a fixed service-capacity model (see section 3). We first discuss the model with immediate access and then come back to the model with delayed access.

6.1 Immediate Access

Denote the customer delay of the k-th arriving customer as d_k, $k \geq 1$. Then the following Lindley-type equation is easily constructed:

$$d_{k+1} = [d_k - t_k]^+ + v_{k+1}, \tag{19}$$

with t_k the number of slots between the arrivals of customers k and $k+1$, v_{k+1} the service time of the $(k+1)$-st customer and $[.]^+$ shorthand for $\max(.,0)$. This is the same (Lindley-type) system equation as in the common discrete-time single-server queue with non-zero service times and thus the pgf $D(z)$ of the delay of a random customer is also the same [1], in this case,

$$D(z) = \frac{1-\lambda\nu}{\lambda} \cdot \frac{(z-1)V(z)\left[A\left(V(z)\right)-1\right]}{\left[z-A\left(V(z)\right)\right]\left[V(z)-1\right]}. \tag{20}$$

Substitution of (4) results in the pgf of the customer delay in the variable service-capacity system:

$$D(z) = \frac{1-\lambda\nu}{\lambda} \cdot \frac{(z-1)S\left(\frac{\mu}{1+\mu-z}\right)\left[A\left(S\left(\frac{\mu}{1+\mu-z}\right)\right)-1\right]}{\left[z-A\left(S\left(\frac{\mu}{1+\mu-z}\right)\right)\right]\left[S\left(\frac{\mu}{1+\mu-z}\right)-1\right]}.$$

6.2 Delayed Access

Now, we assume that arrivals occur after departures at a slot boundary, leading to a minimal delay of 1 slot for each customer. The system equation for the customer delay then reads

$$\tilde{d}_{k+1} = \begin{cases} \tilde{d}_k - t_k + v_{k+1} & \text{if } \tilde{d}_k > t_k \\ 1 + v_{k+1} & \text{if } \tilde{d}_k \le t_k \end{cases}, \tag{21}$$

with $\tilde{d}_k$ the delay of customer k in this model with delayed access, and the other variables as defined in the former subsection. The difference with the Lindley-equation (19) is the term '1' on the second line. We transform the system equation (21) as follows

$$\tilde{d}_{k+1} - 1 = \begin{cases} (\tilde{d}_k - 1) - t_k + v_{k+1} & \text{if } \tilde{d}_k > t_k \\ v_{k+1} & \text{if } \tilde{d}_k \le t_k \end{cases},$$

which can be rewritten as

$$(\tilde{d}_{k+1} - 1) = [(\tilde{d}_k - 1) - t_k]^+ + v_{k+1}.$$

From this equation and (19), it is easily seen that $\{d_k\}_{k\ge1}$ and $\{\tilde{d}_k - 1\}_{k\ge1}$ obey the same system equations and thus

$$D(z) = \frac{\tilde{D}(z)}{z},$$

with $\tilde{D}(z)$ the pgf of the customer delay in the model with delayed access. Using (20), we finally retrieve

$$\tilde{D}(z) = \frac{1-\lambda\nu}{\lambda} \cdot \frac{z(z-1)V(z)\left[A\left(V(z)\right)-1\right]}{\left[z-A\left(V(z)\right)\right]\left[V(z)-1\right]}.$$

Substitution of (4) results in the pgf of the customer delay in the variable service-capacity system:

$$\tilde{D}(z) = \frac{1-\lambda\nu}{\lambda} \cdot \frac{z(z-1)S\left(\frac{\mu}{1+\mu-z}\right)\left[A\left(S\left(\frac{\mu}{1+\mu-z}\right)\right)-1\right]}{\left[z-A\left(S\left(\frac{\mu}{1+\mu-z}\right)\right)\right]\left[S\left(\frac{\mu}{1+\mu-z}\right)-1\right]},$$

which yields the expression obtained in [5].

7 Conclusions

In this paper, we studied a discrete-time queue with variable (geometric) service capacity. The analysis is heavily based on inventive stochastic arguments rather than heavy algebraic calculations. Indeed, we showed how this system is identical to a fixed service-capacity system with non-negative service times. The buffer occupancy was analyzed by connecting the buffer occupancy as seen by departing customers, the buffer occupancy as seen by arriving customers and the buffer occupancy at slot boundaries. The natural single-server model with non-negative service times is a late arrival system with immediate access, and therefore we studied this first. However, service requirements in the original model are strictly positive and a late arrival system with delayed access is more natural in that context. Therefore, we extended the fixed service-capacity analysis to this system. In future work, we plan on tackling non-geometric service capacities. However, the service times in the analog single-server system are then correlated, which is expected to greatly complicate the analysis.

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

References

1. Bruneel, H., Kim, B.: Discrete-time models for communication systems including ATM. Kluwer Academic Publisher, Boston (1993)
2. Bruneel, H.: Performance of discrete-time queueing systems. Computers and Operations Research **20** (1993) 303–320
3. Gao, P., Wittevrongel, S., Bruneel, H.: Discrete-time multiserver queues with geometric service times. Computers and Operations Research **31** (2004) 81–99
4. Janssen, A., van Leeuwaarden, J.: Analytic computation schemes for the discrete-time bulk service queue. Queueing Systems **50** (2005) 141–163
5. Bruneel, H., Wittevrongel, S., Claeys, D., Walraevens, J.: Analysis of a discrete-time queue with geometrically distributed service capacities. In: Proceedings of the 19th International Conference on Analytic and Stochastic Modelling Techniques and Applications (ASMTA12), LNCS 7314, Grenoble (2012) 121–135

6. Kafetzakis, E., Kontovasilis, K., Stavrakakis, I.: 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
7. Chang, C., Thomas, J.: Effective bandwidth in high-speed digital networks. IEEE Journal on Selected Areas in Communications **13** (1995) 1091–1100
8. Jin, X., Min, G., Velentzas, S.: An analytical queuing model for long range dependent arrivals and variable service capacity. In: Proceedings of IEEE International Conference on Communications (ICC 2008), Beijing (2008) 230–234
9. Glock, C.: Batch sizing with controllable production rates. International Journal of Production Research **48** (2010) 5925–5942
10. Balkhi, Z.: On the global optimal solution to an integrated inventory system with general time varying demand, production and deterioration rates. European Journal of Operations Research **114** (1999) 29–37
11. Yang, H.L.: A partial backlogging production-inventory lot-size model for deteriorating items with time-varying production and demand rate over a finite time horizon. International Journal of Systems Science **42** (2011) 1397–1407
12. Hunter, J.: Mathematical Techniques of Applied Probability, Volume 2, Discrete Time Models: Techniques and Applications. Academic Press, New York (1983)
13. Takagi, H.: Queueing analysis: a foundation of performance evaluation, volume 3: discrete-time systems. North-Holland (1991)
14. Fiems, D., Steyaert, B., Bruneel, H.: Discrete-time queues with generally distributed service times and renewal-type server interruptions. Performance Evaluation **55** (2004) 277–298
15. Takagi, H.: Queueing analysis: a foundation of performance evaluation, volume 1: vacation and priority systems, part 1. North-Holland (1991)
16. Bruneel, H., Steyaert, B., Desmet, E., Petit, G.: An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues. International Journal of Digital and Analog Communication Systems **5** (1992) 193–201
17. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press (2008)