Analysis of a discrete-time GI–G–1 queueing model subjected to bursty interruptions

Dieter Fiems *, Bart Steyaert, Herwig Bruneel

SMACS Research Group, Vakgroep TELIN (TW07V), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

Received 1 February 2001; accepted 1 July 2001

Abstract

In this contribution, we investigate a discrete-time single-server queue subjected to server interruptions generated by a 2-state Markov process. The model under consideration assumes customers with multiple-slot service times, which leads to the introduction of two different service strategies depending on whether service of an interrupted customer continues or restarts after an interruption. For both alternatives, we establish expressions for the steady-state probability generating functions of the buffer contents, the unfinished work and the customer delay in terms of the effective customer service times. From these results, closed-form expressions for various performance measures, such as the moments of these quantities, can be established. After dealing with some stability issues, we illustrate the impact of both service strategies on the buffer performance with some numerical examples.

Scope and purpose

Discrete-time queueing theory is distinguished from its more developed continuous-time counterpart by the synchronous nature of its service. Time is divided into fixed-length slots and service of a customer typically takes an integer number of slots. Discrete-time queueing models are particularly appropriate to describe the various queueing related phenomena in digital computer and communication systems including mobile and B-ISDN networks based on asynchronous transfer mode (ATM) technology, due to the packetized nature of these transport protocols.

This work focuses on a discrete-time queueing system with server interruptions. Interruptions are an abstraction of temporary server unavailability. The latter can be caused by temporary occupancy of the server by other customers — regulated by a scheduling discipline such as round robin or priority scheduling — or temporary failure of the server (e.g. due to transmission errors, maintenance or repair). Typically, interruptions occurring in these systems have a bursty nature, i.e., they occur in bursts rather than being randomly spread over time. The presented model allows to study the effects of the burstiness of the server interruptions on various performance measures such as mean queue contents and mean

* Corresponding author. Tel.: +32-9-264-89-02; fax: +32-9-264-42-95.

E-mail address: dieter.fiems@telin.rug.ac.be (D. Fiems).
customer delay. Two different operation strategies are hereby considered: after an interruption, service may either continue where it left off or it may completely restart. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Discrete-time queueing models; Service strategies; Markovian server interruptions

1. Introduction

Queueing models with server interruptions [1] have been investigated by several authors, in continuous as well as in discrete time. Server interruptions are an abstraction of temporary server unavailability caused by sharing a common server with other queues — e.g. for polling systems [2] or multi-class queueing systems — or by external causes, e.g. for maintenance [3] or due to failures or interference [4], etc.

The current contribution investigates a discrete-time single-server queueing model with correlated server interruptions, in the sense that the interruption process is a two-state Markov process. The length of the customer service-times is assumed to be generally distributed. Such a model can be used to assess the performance, for instance, of an IP-based network, where variable-length packets are exchanged between network nodes. The core contribution of this paper is that it takes the possible occurrence of interruptions explicitly into account. We will investigate the cases where after an interruption, service of a customer either continues or restarts. In case the interruptions are caused by servicing higher priority customers in a preemptive priority queue, these two modes correspond to preemptive resume and the preemptive non-resume (without resampling) service disciplines.

Using a generating-functions approach, we present an analysis that allows us to derive expressions for the moments of quantities such as the buffer contents, the unfinished work and the customer delay. In particular, the effect of correlation and of the different service strategies on system performance is investigated. The model extends the results presented in [5,6] as well as those presented in [7,8], in that it considers, respectively, correlated server interruptions (compared with no interruptions in [5] and uncorrelated random interruptions in [6]) and generally distributed service times (compared with single slot service times in [7] and fixed-length service times in [8]).

The outline is as follows. In the next section, the system under investigation is described in more detail. Expressions for the probability generating function of the effective service time are derived in Section 3. The results of the latter are used in Sections 4 and 5, where we derive expressions for the probability generating function of the buffer occupancy and of the unfinished work and message delay, respectively. After some remarks on stability in Section 6 and a numerical example in Section 7, we conclude in Section 8.

2. Model

We consider a discrete-time system, i.e., time is divided into constant-length intervals called slots. During each of these slots, customers that arrive in the system are stored in a buffer
with infinite capacity, and are served on a first-in-first-out basis. Service of a customer takes, in general, a number of slots and is synchronized with respect to slot boundaries. This implies that the service of a customer cannot start before the beginning of the slot following its arrival slot. The numbers of customer arrivals during the consecutive slots are modeled as a series of i.i.d. non-negative random variables with common probability mass function \( e(k) (k \geq 0) \), and common probability generating function \( E(z) \),

\[
E(z) = \sum_{k=0}^{\infty} e(k)z^k.
\]

The service times of the consecutive customers — which is the number of slots the customer would require for service in an environment without server interruptions — are modeled as a series of i.i.d. positive random variables with common probability mass function \( t(k) (k > 0) \), and a corresponding common probability generating function \( T(z) \),

\[
T(z) = \sum_{k=1}^{\infty} t(k)z^k.
\]

Due to the external causes, the server is not always available. The server alternates between an available (A) state and an interrupted or blocked (B) state. The durations of the consecutive A and B times (a period during which the output remains in the A and B state, respectively), expressed in slots, are modeled as two independent series of i.i.d. positive random variables with common probability generating functions \( A(z) \) and \( B(z) \), respectively. Whenever a B-period occurs, the ongoing transmission of a message (if any) is interrupted. We assume that the A- and B-times are geometrically distributed, i.e.,

\[
A(z) = \frac{(1 - \alpha)z}{1 - \alpha z},
\]

\[
B(z) = \frac{(1 - \beta)z}{1 - \beta z},
\]

where \( \alpha \) and \( \beta \) denote the probabilities that the server remains in the A- or B-state, respectively. Therefore, the output process can also be considered as being a correlated two-state Markovian process with states A and B, and with transition probabilities at slot boundaries given by \((1 - \alpha)\) and \((1 - \beta)\) from A to B and from B to A, respectively.

It is often convenient to use the fraction of available-slots \( \sigma \) and the interruption burstiness factor \( K \) instead of the distribution parameters \( \alpha \) and \( \beta \), i.e.,

\[
\sigma = \frac{1 - \beta}{2 - \alpha - \beta},
\]

\[
K = \frac{1}{2 - \alpha - \beta}.
\]

Given the fraction \( \sigma \), the parameter \( K \) takes values between \( \max(\sigma, 1 - \sigma) \) and infinity and is a measure for the absolute lengths of the A- and B-times. The output burstiness factor \( K \) equals 1 for uncorrelated interruptions, in which case the server is available with probability \( \sigma \) during each slot.
Throughout this paper, we consider two service modes. In the continue after interruption (CAI) mode, service of an interrupted customer resumes after the interruption, as opposed to the repeat after interruption (RAI) mode, where service restarts from the beginning. The superscripts CAI and RAI are used to differentiate between both modes whenever necessary.

3. Effective service times

The effective service time of a customer is defined as the time period — expressed as an integer number of slots — that the system effectively spends on serving this customer. This includes the time during which the server is blocked when the customer is in service and, for RAI, the time lost by restarting the customer’s service from the beginning after an interruption. Due to the independence of the service times of the consecutive customers and the nature of the interruption process described above, the effective service times for the consecutive customers given the server state during the slot preceding the effective service, constitute a series of independent positive random variables, whose distributions depend on the service strategy under consideration and on the state of the server (A or B) during the slot preceding the start of the service.

The CAI mode is a memoryless strategy in the sense that during the effective service no record of the complete service time needs to be kept. From a system’s point of view, serving the remainder of a customer’s service time is equal to serving a new customer with service time equal to the remaining service time of the former customer. This implies that the state of the server during the slot preceding the remaining effective service time and the remaining service time completely determine the remaining effective service time distribution. These remarks yield the following recursive equations by conditioning on the state of the server during the first slot of the effective service time,

\[
s_{k,A}(n) = z s_{k-1,A}(n - 1) + (1 - z) s_{k,B}(n - 1) \quad \text{for } k > 1, \ n > 0,
\]

\[
s_{k,B}(n) = (1 - \beta) s_{k-1,A}(n - 1) + \beta s_{k,B}(n - 1) \quad \text{for } k > 1, \ n > 0,
\]

\[
s_{k,A}(0) = s_{k,B}(0) = 0 \quad \text{for } k > 0,
\]

where \( s_{k,A}(n) \) and \( s_{k,B}(n) \) denote the probability that the effective service time of a customer with service time equal to \( k \) equals \( n \), given the slot preceding the start of service is an A- or B-slot, respectively. Taking the \( z \)-transform of these equations leads to

\[
S_{k,A}(z) = \frac{\alpha + (1 - \alpha - \beta) z}{1 - \beta z} S_{k-1,A}(z),
\]

\[
S_{k,B}(z) = \frac{1 - \beta}{\alpha + (1 - \alpha - \beta) z} S_{k,A}(z)
\]

for \( k > 1 \) with \( S_{k,A}(z) \) and \( S_{k,B}(z) \) the (conditional) probability generating functions corresponding with \( s_{k,A}(n) \) and \( s_{k,B}(n) \), respectively. One can easily verify that (4) remains valid
for \( k = 1 \) by defining \( S_{0,A}(z) = 1 \), i.e., a customer with service time equal to zero immediately leaves the buffer with probability 1.

Solving these recursive equations and summation of these (conditional) probability generating functions over all the possible values of \( k \) with respect to the service time distribution then yields expressions for the probability generating functions of the effective customer service times for CAI, given the state of the server during the slot preceding the start of its effective service:

\[
S^\text{CAI}_A(z) = T \left( \frac{x + (1-x-\beta)z}{1-\beta z} \right),
\]

\[
S^\text{CAI}_B(z) = \frac{1-\beta}{x + (1-x-\beta)z} S^\text{CAI}_A(z).
\]

In case of RAI, we have to take into account that service completely restarts after an interruption. We therefore define \( s^\text{RAI}_{k,A}(n) \) as the probability that the remaining effective service time of a customer with service time equal to \( k \) equals \( n \) in case of RAI, given that the remaining service time of the customer equals \( l \) and given that the slot preceding the remaining service is an A-slot. Similar to the case of CAI, \( s^\text{RAI}_{k,B}(n) \) denotes the probability that the remaining effective service time of a customer with service time equal to \( k \) equals \( n \), given that the slot preceding the remaining service time is a B-slot. Note that in the latter case, the remaining service time equals the service time, as the customer’s service starts from the beginning after a B-slot. Conditioning on the state of the server during the first slot of the (remaining) effective service time yields recursive expressions for these probabilities,

\[
s^\text{RAI}_{k,A}(n) = \alpha s^\text{RAI}_{k,l-1,A}(n-1) + (1-\alpha) s^\text{RAI}_{k,B}(n-1) \quad \text{for } k \geq l > 0, n > 0,
\]

\[
s^\text{RAI}_{l,B}(n) = \beta s^\text{RAI}_{l,l-1,A}(n-1) + (1-\beta) s^\text{RAI}_{l,B}(n-1) \quad \text{for } l > 0, n > 0,
\]

\[
s^\text{RAI}_{k,A}(0) = s^\text{RAI}_{k,B}(0) = 0 \quad \text{for } k \geq l > 0.
\]

Let \( S^\text{RAI}_{k,A}(z) \) and \( S^\text{RAI}_{l,B}(z) \) denote the probability generating functions of \( s^\text{RAI}_{k,A}(n) \) and \( s^\text{RAI}_{l,B}(n) \), respectively, then (7) transforms into

\[
S^\text{RAI}_{k,A}(z) = z S^\text{RAI}_{k,l-1,A}(z) + (1-z) S^\text{RAI}_{k,B}(z),
\]

\[
S^\text{RAI}_{l,B}(z) = \frac{(1-\beta)z}{1-\beta z} S^\text{RAI}_{l,l-1,A}(z)
\]

for \( k \geq l > 1 \). One can easily verify that (8) remains valid for \( k \geq l = 1 \) by defining \( S^\text{RAI}_{k,0,A}(z) = 1 \) for all \( k \) — i.e., the remaining effective service time equals zero slots with probability 1 given that the remaining service time equals zero slots. Equation (8) then easily yields explicit expressions for both \( S^\text{RAI}_{k,A}(z) \) and \( S^\text{RAI}_{l,B}(z) \).

The (conditional) probability generating function of the effective service time of a customer with service time equal to \( k \) then equals \( S_{k,A}(z) \) or \( S_{k,B}(z) \) depending on the state of the server during the slot preceding service. Summation of these (conditional) probability generating
functions over all the possible values of $k$ with respect to the service time distribution then yields expressions for the probability generating functions of the effective service times for RAI, given the state of the server during the slot preceding the start of its service:

$$S^\text{RAI}_A(z) = \sum_{k=1}^{\infty} t(k) S^\text{RAI}_{k,k,A}(z),$$

$$S^\text{RAI}_B(z) = \frac{1 - \beta}{\alpha + (1 - \alpha - \beta)z} S^\text{RAI}_A(z)$$

with,

$$S^\text{RAI}_{k,k,A}(z) = \frac{\alpha^k z^k (1 - \alpha z)(\alpha + (1 - \alpha - \beta)z)}{\alpha(1 + (1 - \alpha - \beta)z)(1 - z) + (1 - \alpha)(1 - \beta)\alpha^k z^{k+1}}.$$  

These expressions are in general not explicit due to the infinite sum on the right-hand side of (9). However, one can easily observe that the values of all derivatives of these probability generating functions evaluated in $z = 1$ — and as a consequence, all moments of the corresponding random variables — can be explicitly calculated in terms of the system parameters (under the assumption that all moments of the service time distribution $t(k)$ exist).

Finally, note that, from (6) and (10), the same relation between $S_A(z)$ and $S_B(z)$ is valid for CAI as well as for RAI. This will enable us to express the results of the next few sections only in terms of $S_A(z)$.

4. Buffer contents

In the first step, we derive an expression for the probability generating function of the buffer occupancy at customer departure times. Let $u_n$ denote the buffer occupancy, i.e., the number of customers in the buffer, at the beginning of the slot following the departure slot of the $n$th customer. For positive $u_n$, service of the $(n+1)$th customer can start immediately. This implies that — as the previous slot was an A slot since there was a departure — it will take $s_A$ slots to the next departure, with $s_A$ as a random variable representing the effective service time of a customer given that its service was preceded by an A-slot, and whose probability generating function is given by (5) or (9) depending on the operation mode under consideration. The buffer occupancy $u_{n+1}$ is then given by

$$u_{n+1} = u_n - 1 + \sum_{j=1}^{s_A} e_j \quad \text{for } u_n > 0$$

with $e_j$ being the number of customers arriving in the system during the $j$th slot of the effective service time of the $(n+1)$th customer.

If, on the other hand, the buffer is empty after the departure of the $n$th customer, then the next customer will not be served immediately. In this case, let $m$ denote the first slot following the departure slot during which one or more customers arrive in the system, and let $e_m$ and $t_m$ denote the number of arrivals and the state of the server during this slot, respectively. As the
service of the \((n+1)\)th customer starts in the slot following slot \(m\) and its effective service time is described by the random variable \(s_{m}\), \(u_{n+1}\) is given by

\[
u_{n+1} = e_m - 1 + \sum_{j=1}^{s_m} e_j \quad \text{for } u_n = 0.
\]

(13)

As the numbers of customer arrivals during consecutive slots constitute a series of i.i.d. random variables, the common probability generating function of the \(e_j\)'s equals \(E(z)\). Furthermore, as the only distinction regarding the number of arrivals between a random slot and the slot \(m\) is that we are certain there arrives at least one customer in the system during slot \(m\), the probability generating function of \(e_m\) is given by

\[
E_m(z) = \frac{E(z) - E(0)}{1 - E(0)}.
\]

(14)

Now define \(q_k\) as the probability that the \(k\)th slot after an A-slot is an A-slot. Explicit expressions for these values are easily derived by means of the following recursive equation and boundary condition:

\[
q_k = \alpha q_{k-1} + (1 - \beta)(1 - q_{k-1}) \quad \text{for } k > 0,
q_0 = 1,
\]

(15)

also leading to an expression for the corresponding generating function \(Q(z) = \sum_{k=1}^{\infty} q_k z^k\),

\[
Q(z) = \frac{\alpha + z(1 - \alpha - \beta)}{1 + z(1 - \alpha - \beta)} \frac{z}{1 - z}.
\]

(16)

Note that \(Q(z)\) is a \(z\)-transform of a time series and not a probability generating function.

Due to the nature of the arrival process, slot \(m\) is the \(k\)th slot after the last departure slot with probability \(E(0)^{k-1}(1 - E(0))\). Furthermore, this slot is an A-slot with probability \(q_k\) as the server is available during the last slot of the effective service time of the preceding customer. Summation over all possible values of \(k\) yields an expression for the probability \(\gamma\) that the server is available during slot \(m\),

\[
\gamma = Q(E(0)) \frac{1 - E(0)}{E(0)}.
\]

(17)

Now assume the existence of a stationary distribution (cf. Section 6), i.e., \(U(z) = U_{n+1}(z) = U_n(z)\) as \(n\) approaches infinity. From (12)–(14) and (17), it then follows that the probability generating function of the buffer occupancy at departure times is given by

\[
U(z) = U(0) \frac{(1 - \gamma E(z) - \tilde{\gamma} E(0))S_A(E(z)) - \tilde{\gamma}(E(z) - E(0))S_B(E(z))}{(1 - E(0))(S_A(E(z)) - z)},
\]

(18)

where \(S_A(z)\) and \(S_B(z)\) are given by (5), (6) or by (9), (10) depending on the service strategy and with \(\tilde{\gamma}\) shorthand for \((1 - \gamma)\). This further simplifies — in view of the remark at the end
of the previous section — to
\[
U(z) = U(0) \frac{E(z)}{E(0)} \frac{Q(E(0))}{Q(E(z))} \frac{S_\Lambda(E(z))}{S_\Lambda(E(z)) - z}.
\] (19)

The unknown parameter \(U(0)\) in (18) and (19) can be determined by applying the normalization condition \(U(1) = 1\), leading to
\[
U(0) = \frac{\sigma E(0)}{Q(E(0))} \frac{1 - \mu E \mu S_\Lambda}{\mu E},
\] (20)
with \(\mu E\) the mean number of customers arriving in a slot, and \(\mu S_\Lambda\) the mean effective service time given that the output is available during the slot preceding the start of the customer’s service. The latter can be determined by differentiating (5) or (9) in case of CAI or RAI, respectively. Combining Eqs. (19) and (20) finally yields the probability generating function of the buffer occupancy at departure times,
\[
U(z) = \sigma \frac{1 - \mu E \mu S_\Lambda}{\mu E} \frac{E(z)}{Q(E(z))} \frac{S_\Lambda(E(z))}{S_\Lambda(E(z)) - z}.
\] (21)

The probability generating function of the buffer occupancy at random slot boundaries \(N(z)\) relates to this probability generating function as follows [5]:
\[
N(z) = \frac{U(z)(z - 1)\mu E}{E(z) - 1}.
\] (22)

Finally, taking the appropriate derivatives of (21) and (22) yields expressions for the various moments of the buffer occupancy at departure times and at random slot boundaries, respectively. These expressions are explicit for CAI as well as for RAI.

5. Unfinished work and customer delay

The unfinished work at a given time instant is defined as the number of slots it would take to empty the buffer under the assumption that there are no new arrivals. Note that this includes the B-slots during which the server is interrupted, and (in case of RAI) slots lost due to uncompleted interrupted service trials. Consider now a random slot \(k\), and define \(w_k\) as the unfinished work at the beginning of this slot. During this slot, the unfinished work is diminished by 1 if the buffer is not empty. For each customer arriving in slot \(k\), an additional number of slots equal to the effective service time of this customer is required to empty the buffer, i.e., the unfinished work at the beginning of slot \((k + 1)\) relates to the unfinished work at the beginning of slot \(k\) as follows:
\[
w_{k+1} = (w_k - 1)^+ + \sum_{j=1}^{\cdot} s_j
\] (23)
with \(\cdot\) the random variable denoting the number of arrivals in slot \(k\), \(s_j\) the random variable denoting the effective service time of the \(j\)th customer arriving in slot \(k\) and \((\cdot)^+\) shorthand for \(\max(0, \cdot)\).
The second term on the right-hand side of Eq. (23) equals zero when there are no arrivals in slot $k$. If there is at least one arrival, then all $s_j$’s have $S_A(z)$ as common probability generating function — service of the next customer immediately starts after the previous one, i.e., after an A-slot — in case the buffer is not empty at the beginning of slot $k$. This is also the case for all $s_j$’s but the first if the buffer is empty. When a new batch of customers arrive in an empty buffer during slot $k$, the service of the first customer starts during slot $k+1$, and it will have $S_A(z)$ or $S_B(z)$ as the corresponding probability generating function depending on the state of the server during slot $k$. One easily proves that customers arriving in an empty system, arrive in an A-slot with probability $\gamma$.

Summarizing, let $W_k(z)$ denote the probability generating function of $w_k$, then,

$$W_{k+1}(z) = W_k(z)E(S_A(z)) \frac{1}{z} + W_k(0) \left( C(z) - \frac{1}{z}E(S_A(z)) \right)$$

with

$$C(z) = E(0) + \left( \gamma + (1 - \gamma) \frac{S_B(z)}{S_A(z)} \right) (E(S_A(z)) - E(0)).$$

Now assume the existence of a stationary distribution, i.e., $W(z) = W_{k+1}(z) = W_k(z)$ for $k$ approaching infinity. Using the fact that an empty buffer implies zero unfinished work and vice versa, i.e., $W(0) = N(0)$, the remark at the end of Section 3, Eq. (24) transforms into

$$W(z) = \frac{\sigma}{\gamma} \frac{zC(z) - E(S_A(z))}{z - E(S_A(z))}. \quad (26)$$

Consider now a particular (tagged) customer. The probability generating function of the unfinished work at the beginning of its arrival slot $k$ is given by (26) due to the i.i.d. nature of the arrival process. Furthermore, the probability generating function of the number of customers $e_b$ that arrive in the same slot and will be served before the customer under consideration, is given by [5]

$$E_b(z) = \frac{E(z) - 1}{(z-1)\mu_E}. \quad (27)$$

As each of these customers need to be served before the tagged customer, the delay $d$ of the tagged customer is the sum of the unfinished work $w_k$ minus 1 if the buffer is non-empty at the beginning of slot $k$, of the total effective service time of the customers arriving in slot $k$ before the tagged customer — i.e., the sum of the effective service times $s_j$ ($j = 1 \ldots e_b$) of the customers arriving in slot $k$ before the tagged customer — and of the effective service time of the tagged customer $s_{e_b+1}$,

$$d = (w_k - 1)^+ + \sum_{j=1}^{e_b+1} s_j. \quad (28)$$

As explained before, the series of random variables $s_j$ have $S_A(z)$ as common probability generating function if the buffer is non-empty at the beginning of slot $k$. This is also the case for all but the first (possibly the tagged) customer if the buffer is empty at the beginning of
slot \( k \). In this case, the first customer has again \( S_A(z) \) or \( S_B(z) \) as probability generating function depending on the state of server during slot \( k \), i.e., with probability \( \gamma \) or \((1-\gamma)\), respectively. The probability generating function \( D(z) \) of the random variable \( d \), i.e., of the delay of a random customer, is then given by

\[
D(z) = \frac{E_b(S_A(z))}{z} (W(z)S_A(z) + W(0)((\gamma z - 1)S_A(z) + (1 - \gamma)zS_B(z))).
\]  

(29)

This result further simplifies to

\[
D(z) = \sigma \frac{1 - \mu E\mu_s}{\mu_E} \frac{z}{Q(z)} \frac{S_A(z)}{S_A(z) - 1} \frac{1 - E(S_A(z))}{z - E(S_A(z))}.
\]  

(30)

The various moments of unfinished work and customer delay can then be determined by taking the appropriate derivatives of (26) and (30), respectively.

6. Stability

In the remainder of the paper we assume non-heavy-tailed customer arrival and service time distributions, i.e., at least the first and second moments of these random variables are finite.

Consider now the effective service times. It is clear from (3), (5) and (6) for CAI and from (7), (9) and (10) for RAI, that as long as \( \alpha > 0 \) and \( \beta < 1 \), the conditional effective service time distributions (conditioned on the state of the server during the slot preceding the start of the customer service) exist and that the corresponding generating functions are given by (5), (6) and (9), (10) for CAI and RAI, respectively. The existence of the distributions, however, does not necessarily guarantee the existence of their moments. For CAI, evaluation of the appropriate derivatives of (5) and (6) in \( z = 1 \) easily shows that the \( n \)th moments of the conditional effective service time distributions are functions of the moments of the service time up to and including order \( n \), and of the service parameters \( \alpha \) and \( \beta \). As a consequence, for CAI the effective service time distributions can only exhibit heavy-tailed behavior inherited from the underlying service time distribution. In particular, due to the initial assumptions, the first and second moments of the conditional effective service times exist in case of CAI.

For RAI, evaluation of the appropriate derivatives of (9) and (10) in \( z = 1 \) yields the dependency of the conditional effective service times’ moments on the complete service time distribution. Define \( R_T \) as the (possibly infinite) radius of convergence of the probability generating function \( T(z) \) of the service time. The \( n \)th moment of the conditional effective service time distributions then exist if \( \alpha > R_T^{-1/n} \) and are infinite if \( \alpha < R_T^{-1/n} \); the case \( \alpha = R_T^{-1/n} \) depends on the behavior of \( T(z) \) and its derivatives at \( z = R_T \).

A necessary and sufficient condition for the existence of a stationary distribution for the buffer occupancy, unfinished work and customer delay is that the effective load \( \rho \) is less than the system capacity, or equivalently, that the probability of having an empty buffer is strictly positive,

\[
\rho = 1 - N(0) < 1.
\]  

(31)
Using (5) in case of CAI and (9) in case of RAI, this inequality reduces to
\[ \mu_T \mu_E < \frac{1 - \beta}{2 - \alpha - \beta} \]  
(32)
and
\[ \frac{\alpha}{1 - \alpha} \left( T \left( \frac{1}{\alpha} \right) - 1 \right) \mu_E < \frac{1 - \beta}{2 - \alpha - \beta} \]  
(33)
respectively, where \( \mu_T \) represents the mean customer service time. Therefore, due to (33), \( \alpha > R_T^{-1} \) is a necessary condition for having a stable system in case of RAI. The existence of a stationary distribution, however, does not guarantee in general finite stationary moments. From (18), (22), (24) and (29), it is obvious that the moments of the buffer occupancy at departure times, the buffer occupancy at random slot boundaries, the unfinished work at random slot boundaries and the customer delay are functions of the moments of the effective service time distributions and of the customer arrival distribution. In particular, the expressions for the means of these random variables contain the first and second moments of the conditional effective service times and the first and second moments of the customer arrival distribution.

Summarizing, due to the initial assumptions, in case of CAI, the mean buffer occupancy, the mean customer delay and the mean unfinished work are finite as long as (33) is satisfied. However for RAI, (33) as well as the condition \( \alpha > 1/\sqrt{R_T} \) must be satisfied. Note that only when the customer service times can become infinitely long, this will pose an extra condition on the system parameters, since in the opposite case \( R_T \) equals infinity. In the remainder, we will refer to these requirements as the stability conditions.

### 7. Numerical example

Let the number of customers arriving in a slot be Poisson distributed with mean \( \lambda \) and assume that the service times for the consecutive customers are shifted geometrically distributed with parameter \( \phi \), i.e.,
\[ E(z) = e^{\lambda(z-1)}, \]  
(34)
\[ T(z) = \frac{(1 - \phi)z}{1 - \phi z}. \]  
(35)

In Fig. 1, the stability conditions for this system for CAI and RAI are plotted versus the interruption parameters \((\sigma, K)\) as defined by (2) for different \( \phi \)-values and for \( \lambda = 0.05 \). Given \( \phi \) and \( \lambda \), the mean buffer occupancy is finite for \((\sigma, K)\) to the right of the corresponding condition curve. Note that for large values of \( K \) the stability conditions for RAI equal those for CAI as nearly no customers are interrupted during service when the mean A-times are large. This also follows from Fig. 2(b), where the mean customer delay is depicted versus the available-fraction \( \sigma \) for different values of \( K \). For increasing \( K \), the corresponding CAI and RAI curves come closer to each other, and for sufficiently large values of \( K \), the buffer performance becomes comparable. From Fig. 1, it is also clear that given \( \sigma, \lambda \) and \( \phi \) satisfying the equilibrium conditions for CAI, stability for RAI determines a lower bound for \( K \).
Fig. 1. Stability conditions for CAI and RAI.

Fig. 2. Mean customer delay ($\phi = 0.8$, $\lambda = 0.05$).

Fig. 2(a) depicts the mean customer delay versus the burstiness factor $K$ for different values of $\sigma$. In case of CAI, an increase of $K$ implies a higher mean customer delay due to the increasing burstiness (i.e., longer A- and B-periods) of the interruptions. However, $K$ does not influence the stability conditions. For RAI, we have the superposition of two effects. On the one hand, longer A-times imply that less customers are interrupted during service and as a consequence a lower mean customer delay. This means that for increasing $K$ the mean customer delay...
delay decreases. On the other hand, increasing values for $K$ also imply increasing burstiness in the interruption process and as a consequence also an increasing mean customer delay. It is clear that the former effect dominates for small values of $K$. The A-periods are very short and as a consequence, customers are frequently interrupted during service. Once $K$ has reached a certain threshold, the number of customers interrupted during their service will be low and a further increase will no longer imply a decreasing mean customer delay. For higher $K$, the burstiness effect will dominate. This explains why the curves for RAI in Fig. 2(a) reach a minimal value. Note that the improvement for low $K$ values also follows from Fig. 2(b) in case of RAI. For low $K$ values, system performance improves. A further increase, however, has the reverse effect. Note in particular how for $K=50$ both operation modes perform nearly identically.

In Fig. 3(a), the optimal values for $K$ in case of RAI, i.e., the $K$ values minimizing the mean customer delay, are plotted versus the mean service time for a fixed arrival rate $\lambda = 0.05$ and for different values of $\sigma$. For longer service times, the optimal $K$ values increase and become infinite for $\mu_T\lambda = \sigma$, i.e., as long as the equilibrium condition for CAI is satisfied, one can find an optimal $K$ value. As CAI always outperforms RAI or by comparing Figs. 1(a) and (b), the latter condition is also required to ensure stability for RAI. Fig. 3(a) then implies that as long as the system is stable, there is an optimal $K$ value that minimizes the mean customer delay.

In Fig. 3(b), the effects of the variance of the service time on the mean customer delay is investigated. For this, we use a weighted sum of two shifted geometrically distributed random variables as service time, i.e.,

$$T(z) = \frac{1}{2} \left( \frac{1 - \phi_1 z}{1 - \phi_1 z} + \frac{1 - \phi_2 z}{1 - \phi_2 z} \right),$$

(36)
and a Poisson customer arrival distribution as defined by (34). Fig. 3(b) depicts the mean
customer delay versus \( K \) for a fixed mean service time \( \mu_T = 5 \), fixed arrival rate \( \lambda = 0.05 \), and
different values for the variance \( \sigma_X^2 \) of the service time distribution. Note that given the mean and
variance of the service time distribution, the unknown parameters \( (\phi_1, \phi_2) \) in (36) are easily
determined. For both modes, a higher variance translates in a higher mean customer delay.
Whereas for CAI, this increase is constant for all \( K \), for RAI, an increase of variance translates
into more stringent stability conditions as the radius of convergence of the given service time
distribution decreases for higher variances. We also observe that the variance of the customer
service time has a severe impact on the system performance, especially for RAI in the region
where interruptions of customers during their service are abundant (i.e., low values of \( K \)).

8. Conclusions

We investigated a discrete-time single-server queueing system with Markovian interruptions
and generally distributed service times, using a probability generating functions approach. We
obtained expressions for the probability generating function of the buffer occupancy, unfinished
work and customer delay for two service modes, called CAI and RAI. These expressions can be
used to obtain closed-form expressions for the moments of the corresponding random variables.
Wherever stability requirements under fairly general conditions are limited to the equilibrium
condition for CAI, we showed that this is not the case for RAI. An additional condition, solely
dependent on the length of the mean on-period, must be satisfied. Furthermore, a comparison
with the uncorrelated interruption model yields the fact that correlation effects are completely
different for CAI and RAI.

When the correlation in the service process, captured by the parameter \( K \), increases, the
system performance deteriorates for CAI, whereas for RAI the system performance improves
when \( K \) is low and deteriorates when \( K \) is high. Although we were able to model correlation
in the interruption process, this is still a fairly simple model. More sophisticated models are
subject to further study.

References

Dieter Fiems received an engineering degree at KAHO-St-Lieven and a post-graduate degree in computer science at Ghent University. He is now a Ph.D. student at the Department of Telecommunications and Information Processing at Ghent university, as a member of the SMACS research group. His main research interests include discrete-time queueing models and stochastic modeling of IP and ATM networks.

Bart Steyaert received the degrees of Licentiate in Physics and Licentiate in Computer Science from Ghent University, Belgium. Since, January 1990, he has been working as a researcher at the SMACS Research Group. His main research interests include discrete-time queueing models, traffic control, and stochastic modeling of ATM and IP networks.

Herwig Bruneel received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science from Ghent University, Belgium. He leads the SMACS Research Group within this department. His main research interests include stochastic modeling of digital communication systems, discrete-time queueing theory, and the study of ARQ protocols. He has published more than 150 papers on these subjects and is the coauthor of the book “Discrete-Time Models for Communication Systems Including ATM”, (besides B.G. Kim).