Analysis of a discrete-time single-server queue with an occasional extra server

Herwig Bruneel, Sabine Wittevrongel*

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

Abstract

We consider a discrete-time queueing system having two distinct servers: one server, the "regular" server, is permanently available, while the second server, referred to as the "extra" server, is only allocated to the system intermittently. Apart from their availability, the two servers are identical, in the sense that the customers have deterministic service times equal to 1 fixed-length time slot each, regardless of the server that processes them. In this paper, we assume that the extra server is available during random "up-periods", whereas it is unavailable during random "down-periods". Up-periods and down-periods occur alternately on the time axis. The up-periods have *geometrically distributed* lengths (expressed in time slots), whereas the distribution of the lengths of the down-periods is *general*, at least in the first instance. Customers enter the system according to a *general* independent arrival process, i.e., the numbers of arrivals during consecutive time slots are i.i.d. random variables with arbitrary distribution.

For this queueing model, we are able to derive closed-form expressions for the steady-state probability generating functions (pgfs) and the expected values of the numbers of customers in the system at various observation epochs, such as the start of an up-period, the start of a down-period and the beginning of an arbitrary time slot. At first sight, these formulas, however, appear to contain an infinite number of unknown constants. One major issue of the mathematical analysis turns out to be the determination of these constants. In the paper, we show that restricting the pgf of the down-periods to be a *rational* function of its argument, brings about the crucial simplification that the original *infinite* number of unknown constants appearing in the formulas can be expressed in terms of a *finite* number of independent unknowns. The latter can then be adequately determined based on the bounded nature of pgfs inside the complex unit disk, and an extensive use of properties of polynomials.

Various special cases, both from the perspective of the arrival distribution and the down-period distribution, are discussed. The results are also illustrated by means of relevant numerical examples.

Possible applications of this type of queueing model are numerous: the extra server could be the regular server of another similar queue, helping whenever an idle period occurs in its own queue; a geometric distribution for these idle times is then a very natural modelling assumption. A typical example would be the situation at the check-in counter at a gate in an airport: the regular server serves customers with a low-fare ticket, while the extra server gives priority to the business-class and first-class customers, but helps checking regular customers, whenever the priority line is empty.

Keywords: queueing, discrete-time, two servers, server interruptions, polynomials

1. Introduction

In this paper, we analyze a discrete-time infinite-waiting-room single-class *two-server* queueing system with an uncorrelated batch arrival process, individual service (as opposed to batch service), constant service times, and *random server interruptions*, i.e., where the number of (available) servers varies stochastically. Our study only aims at the characterization of the number of customers in the system, i.e., the so-called "system content", and disregards the determination of customer-delay characteristics, which implies that the applied queueing discipline is largely irrelevant. From the point of view of mathematical analysis, the two major difficulties in our model are its *multi-server* character and the presence of *server interruptions*. Some general context on these issues and a summary of related earlier work are given below.

1.1. Multi-server queueing models

As opposed to their single-server counterparts, multi-server queues are notoriously hard to analyze mathematically, unless severe model restrictions are introduced. So far, no explicit analytic results are available for multi-server queues with completely arbitrarily distributed service times, neither in a continuous-time setting nor in a discrete-time setting. Even for the seemingly simple special case where the number of servers is equal to two, no general solution techniques seem to exist.

1.1.1. Continuous-time models

Some papers dealing with *continuous-time* two-server queues with general service-time distribution are (in chronological order) [1, 2, 3, 4, 5]. In [1], an $M/G/2$ model is studied; the joint distribution of the system content and of the remaining holding times for services in progress turns out to be impossible to compute in a general setting; only when the service time has a rational Laplace-Stieltjes transform, a method to obtain the queue-length distribution is constructed. In [2], again, an $M/G/2$ queue is considered, in this case under the restriction that the service times are distributed according to a mixture of (negative) exponential distributions, i.e., a hyperexponential distribution; the numerical complexity of the analysis in [2] is lower than in the case of [1]. The same model as in [1] is studied in [3]; here the problem of determining the marginal distribution of the system content is reduced to the solution of a pair of coupled integral equations. When the two servers are identical, only a single integral equation must be solved. The explicit solution of the integral equation(s) is only achieved for several specific service-time distributions (e.g., Erlang, hyperexponential, and deterministic). Similarly, [4] considers an optimization problem with a single queue and two heterogeneous servers, where the question is when to use only the fast server, only the slow server or both;

*Corresponding author
Email addresses: Herwig.Bruneel@UGent.be (Herwig Bruneel), Sabine.Wittevrongel@UGent.be (Sabine Wittevrongel)

here, again, results are only obtained for specific service-time distributions (exponential, Erlang). The paper [5] considers a heterogeneous M/G/2 queue, where the service times at server 1 are exponentially distributed, and at server 2 they have a rational Laplace-Stieltjes transform, i.e., again results are only obtained under certain restrictions on the service-time distributions.

Continuous-time models with more than two servers and "generalized" service times have also been investigated, e.g., in the papers (in chronological order) [6, 7, 8, 9, 10, 11]. Here, [6] presents an approximate analysis of the waiting-time distribution in the $M/G/c$ queue; the result is given in the form of an integral equation which can only be solved numerically. Paper [7] derives explicit results for the stationary distributions of waiting times and queue lengths in $GI/G/c$ queues, in case the service-time distribution is hyperexponential. The basis of the analysis is the reduction of the problem to the solution of a system of Wiener-Hopf-type equations. In [8], an analytic technique is developed to analyze $GI/G/c$ queues, under the restriction that both the interarrival-time distribution and the service-time distribution are Coxian distributions (which compose a subset of the distributions with rational Laplace-Stieltjes transform). On the other hand, [9] considers a queue with multiple servers, with so-called "generalized exponential" service times, where both the arrival and service processes are modulated by the same finite-state Markov chain. In [10], the authors explicitly comment on the mathematical difficulty of analyzing multi-server queues with completely general service-time distributions. They then opt for the approach to replace the general service-time distribution by a phase-type distribution, motivated by the fact that the $M/PH/c/N$ queue can – in principle – be analyzed by solving a set of linear balance equations for the state probabilities. In order to circumvent the issue of state-space explosion, they propose to use a reduced state description in which the state of only one server is represented explicitly, while the other servers are accounted for through their rate of completions. In [11], a multi-server queue with phase-type distributed service times is considered in which servers are activated or switched off depending on the relation between the queue length and some predefined thresholds.

In many papers studying more complicated issues in the context of continuous-time multi-server queueing models, the most simple assumption of exponential service times is maintained, so as not to complicate the analysis from the point of view of the service process. Some examples are (in chronological order) [12, 13, 15, 14, 16]. Here, [12] studies the transient behavior of a Markov-modulated Poisson arrival queue with multiple exponential servers under overload control. In [13], a $MAP/M/2$ system with two classes of customers is considered, where one type of customers requires only one server and the other type needs both servers. Papers [14, 16] deal with a two-class two-server queue where both customer classes have their own dedicated server and are accommodated in a single FCFS queue; in both papers, the analysis explicitly relies on the memoryless (exponential) nature of the service times, but the mix of both customer classes in the arrival stream is described differently. Similarly, [15] considers a two-class two-server retrial queueing system, where the service times of each class of customers are assumed to be exponentially distributed with class-dependent service rates.

1.1.2. Discrete-time models

The most simple *discrete-time* multi-server queueing models are those where the service times are deterministically equal to 1 time slot. In this subsection we only review

a number of studies of such queues without server interruptions; models with server interruptions are included in subsection 1.2.2.

A discrete-time queueing model with general independent arrival process, $c \geq 1$ identical servers, and constant 1-slot service times is considered, for instance, in [17, 18, 19], where [17] focuses on the system content, [18] discusses the analysis of the customer delay, and [19] develops accurate approximative closed-form expressions for the tail probabilities of both the system content and the delay. The same system but with the general independent arrival process replaced by a general (correlated) discrete-time Markovian batch arrival process, where the batch-size distribution of the arrivals in successive slots is governed by a finite-state Markov chain, has been considered in [20] for an infinite-waiting-room system, and in [21] for the (more difficult) case of a finite-storage-capacity system. In [22], for any discrete-time c-server queue with constant 1-slot service times, a general relationship between the mass functions of the customer delay and the system content is derived under the most general conditions possible with respect to the nature of the arrival process.

Extensions to more general service-time distributions include [23, 24, 25, 26, 27, 28, 29, 30]. In [23, 24], arbitrary-length constant service times (i.e., m-slot service times ($m \geq 1$) instead of 1-slot service times) are considered. Specifically, the system-content distribution is studied in [23], while a relationship between the probability distributions of the system content and the customer delay is established in [24]. By means of this relation, an explicit expression for the probability generating function (pgf) of the delay is obtained from the known pgf of the system content, derived in [23]. The same relationship is also established and more applications of it are explored in [25]. An extension to a 2-state Markovian correlated arrival process is considered in [26]. Geometrically distributed service times have been similarly investigated in [27, 28, 29]: system content analysis in [27], relationship between the pgfs of customer delay and system content in [28], extension to correlated arrival processes in [29]. For the case of independent arrivals, [30] adds the extra complication of customers with balking behavior, i.e., if all the servers are busy, an arriving customer either enters with some probability or balks with the complementary probability. One of the most general discrete-time multi-server models available today is discussed in [31], where Markovian arrivals are combined with a phase-type distribution for the service times. Here, again, the authors emphasize the great (numerical) complexity associated with an analysis of the system.

1.2. Queueing models with random server interruptions

Queues with random server interruptions have also received considerable attention in the queueing literature. A good recent introductory survey of such studies, both in a continuous-time setting and a discrete-time setting, can be found in [32]. A more extensive overview, tailor-made for the present paper, is given below.

1.2.1. Continuous-time models

Some early *continuous-time* infinite-capacity models were reported in [33, 34] for the single-server case and in [35] for the multi-server case. In [33, 34], server interruptions (of the single server) are generated by a Poisson process and have arbitrarily distributed lengths. In other words, the distributions of the available and interrupted periods of the server are exponential and general, respectively. Service times, on the other hand,

have a general distribution. In [33], the distributions of queue length, waiting time, and busy-period duration are characterized by transforms and by moments; [34] studies the virtual-waiting-time process. In [35], the number of servers can be more than one, but all the relevant probability distributions (service times, available periods, interrupted periods of the servers) are exponential. The main result is a closed-form expression for the pgf of the system content in case the number of servers is 1 or 2; for larger numbers of servers, a numerical procedure is established. Two very interesting continuous-time contributions with respect to the single-server case are [36, 37], which present approximate results under quite general model assumptions. Although there are numerous other instances of such continuous-time analyses to be found, we limit ourselves to these examples because our current paper focuses on discrete-time models (with infinite storage room). Nevertheless, these five examples are "typical" in the sense that the multi-server analysis [35] explicitly requires the assumption that the service-time distribution is memoryless, i.e., exponential; the single-server cases investigated in [33, 34, 36, 37], on the contrary, do not require this restriction and can cope with arbitrarily distributed service times.

A specific subclass of server-interruption models contains those where one or more servers are "removable". In such models, the process that determines the number of available servers can still be labelled as "random", but instead of being an external process, it is triggered by the (internal) state of the queueing system itself. Some papers dealing with such systems are (in chronological order) [38, 39, 40, 41, 42, 43, 44], as well as the aforementioned paper [11]. Specifically, [39, 40] study a K-server system ($K > 1$) where an extra server is added at epochs when the number of customers in the system exceeds a forward threshold value, and a server is removed at epochs when the number of customers in the system becomes less than a reverse threshold value. The same idea had been explored before in the context of tandem queueing models for production lines (with finite intermediate buffers) in [38]. A similar model with a number of permanent servers and a number of removable servers has been thoroughly analyzed in [41] with the additional restriction on departing customers known as "nopassing", according to which the customers have to leave the system in the chronological order of their arrival. In [42], a so-called "congestion-based staffing" policy is explored where the number of servers (in this case, inspection booths in border-crossing stations) is adjusted according to the queue length during a planning period, the primary objective being to maintain the average queue length within a certain range. A similar application is reported in [43] in the context of convenience stores, where additional checkout registers (i.e., servers) are opened when the number of customers standing in line exceeds a specified number, in order to reduce lost sales due to balking of newly arriving customers who are unwilling to accept long delays. The recent book chapter [44] considers a 2-server model in which one server is permanently available and the additional server is activated only if the queue length exceeds some fixed preassigned threshold; the major difficulty of the model is that service times have a phase-type distribution. The paper [11], mentioned before, is in many ways a generalization of [44] from two to more than two servers.

1.2.2. Discrete-time models

Let us concentrate now on *discrete-time* server-interruption models with infinite waiting room. Among the first papers to treat such systems were [45, 46] for the case of one single server, and [47] for the multi-server case (with $m > 1$ servers). In these models, the

service times of the customers are assumed to be fixed and equal to exactly one discrete time slot, and the server interruptions are modelled by means of one single parameter σ which indicates the probability that the server is available (in [45, 46]) or all m servers are available (in [47]) during a slot; furthermore, server availability is assumed to be independent from slot to slot. Various extensions of the basic single-server model dealt with in [45, 46] have been reported in (in chronological order) [48, 49, 50, 51, 52, 53, 54, 55]. In all these papers, the service times are still fixed to one slot each, but the nature of the service-interruption process is different. Specifically, the server-interruption process is modelled by an alternating sequence of geometrically distributed (i.e., memoryless) on-periods and off-periods in [48, 53], i.e., the process is controlled by a 2-state Markov chain. This is also the case in [55], with the slight modification that even during on-periods the server is only available with a probability that may be less than 1. A similar model is considered in [54], where the number of states of the underlying Markov chain is an arbitrary (finite) integer, larger than 1. A model with geometric on-periods and general (i.e., arbitrarily distributed) off-periods is analyzed in [49, 50]; a further generalization to on-periods distributed according to a mixture of a finite number of geometric distributions, and still general off-periods, is reported in [51]. This is further extended to on-periods with a rational pgf and general off-periods in [52]. An extension of the basic multi-server model treated in [47] is studied in [56, 57, 58]), where the number of available servers still changes independently from slot to slot but can take any value between 0 and m, i.e., where the interruptions of the m servers do not necessarily occur simultaneously. Specifically, for this "uncorrelated" multi-server interruption model, the system content is analyzed in [56] and the customer-delay is studied in [57], both under the assumption of uncorrelated arrivals from slot to slot, whereas a general relationship between the pgfs of system content and customer delay, valid for any kind of arrival process, is established in [58]. Further extensions of the "uncorrelated" multi-server interruption process studied in [56, 57, 58] into a time-correlated process are considered in [59, 60]. Specifically, [59] models the number of available servers as a first-order Markov chain with state space $\{0, 1, 2, \cdots, m\}$, whereas [60] presents a hybrid model, where arbitrarily distributed "blocked periods", during which none of the m servers is available, alternate on the time axis with geometrically distributed "available periods", during which the number of available servers takes a random value on the set $\{1, 2, \cdots, m\}$ and changes independently from slot to slot.

Various studies have also allowed for more general service-time distributions than the deterministic 1-slot case. In [61], server interruptions (of the single server) are (again) modelled as an on/off process with geometrically distributed on-periods and generally distributed off-periods, but the distribution of the service times is completely arbitrary; the arrivals are uncorrelated from slot to slot. The model in [62] is a variant of this, where the service-time distribution is general and both the arrival process and the server-interruption process are dependent on a common underlying finite-state Markov chain. Also relevant in this context is [63], where sufficient conditions for system stability, both in the single-server and the multi-server case, are established under renewal assumptions for arrival, service and interruption processes.

1.3. Situation of this paper

The current paper is closely related to some of the earlier works mentioned in the previous subsections, in that it considers a multi-server queue (specifically, a queue with

two servers), where one of the servers is available on a permanent basis and the other server is subject to random interruptions. Just as in queueing models with "removable" servers, as discussed above, our model thus includes a server which is only intermittently available. However, in our model, the presence or absence of the additional server is not triggered by the system content exceeding or falling below certain thresholds, but rather is the result of an external random server-interruption process, independent of the system state. The main purpose of our paper is to analyze a model for the interruption process of the additional server that is more general – in terms of the probability distributions of the on-periods and the off-periods – than the models currently available in literature. Our literature review of earlier related work has clearly illustrated that the exact nature of these distributions turned out to be crucial for the complexity (and even the mere feasibility) of the mathematical analysis. In order not to complicate things further, we make the simplest possible assumption with respect to the nature of the service-time distribution, i.e., we assume that service times are deterministically equal to one slot. The details of the model are discussed in section 2.

2. Mathematical model

In this paper, we investigate a discrete-time queueing model with infinite waiting room and two servers. The first server, in the sequel referred to as the "regular" server, is permanently available, while the second server, referred to as the "extra" server, is only allocated to the system intermittently. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as (time) slots. 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. We assume that the service of each customer requires exactly one slot, regardless of whether the regular server or the extra server handles it.

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

$$\begin{aligned} c(n) &\triangleq \text{Prob}[\ n \text{ customer arrivals in one slot }]\ , \quad n \geq 0\ , \\ C(z) &\triangleq \sum_{n=0}^{\infty} c(n)\, z^n\ . \end{aligned} \tag{1}$$

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

$$\lambda \triangleq \sum_{n=1}^{\infty} nc(n) = C'(1)\ . \tag{2}$$

As the extra server is not permanently available, our model basically divides the time axis into two types of time slots, referred to as "up-slots" and "down-slots", respectively. During up-slots, the extra server is available for the service of customers of the considered system; during down-slots, it is not. A sequence of consecutive up-slots between two consecutive down-slots, is named an "up-period" in the sequel. Likewise, the term

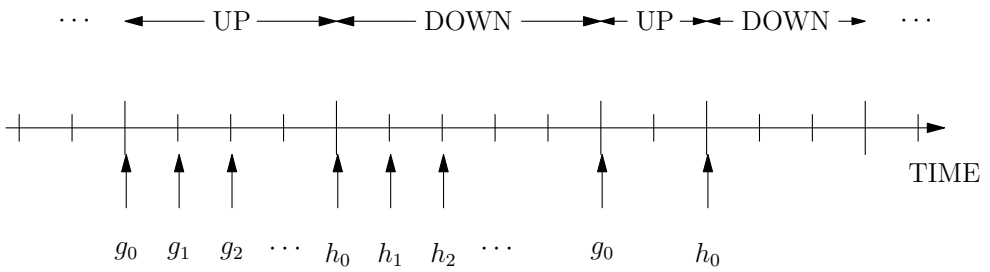

Figure 1: Alternating sequence of up-periods and down-periods. Definition of the random variables $\{g_k\}$ and $\{h_k\}$ for all $k \geq 0$.

"down-period" refers to a sequence of down-slots in between two consecutive up-slots. Up-periods and down-periods occur alternately as time goes by. They are illustrated graphically in Fig. 1. We assume that the lengths of the up-periods are i.i.d. discrete random variables with a common geometric distribution with parameter α ($0 \leq \alpha < 1$). The pmf of an up-period is given explicitly by

$$\text{Prob}[\text{ up-period} = n \text{ slots }] = (1-\alpha)\alpha^{n-1} \quad , \quad n \geq 1 \ ; \tag{3}$$

the mean length of an up-period is given by

$$E[\text{up-period}] = \frac{1}{1-\alpha} \geq 1 \ . \tag{4}$$

The lengths of the consecutive down-periods are also modelled as a set of i.i.d. random variables. In a first instance, we make no restricting assumptions as to the specific nature of their distribution. Their pmf, pgf and mean value are indicated as $r(n)$, $R(z)$ and $\bar{r}$, respectively, i.e.,

$$\begin{aligned} &\text{Prob}[\text{ down-period} = n \text{ slots }] = r(n) \quad , \quad n \geq 1 \ ; \\ &R(z) \triangleq \sum_{n=1}^{\infty} r(n)\, z^n \ ; \\ &\bar{r} \triangleq E[\text{down-period}] = \sum_{n=1}^{\infty} n r(n) = R'(1) \geq 1 \ . \end{aligned} \tag{5}$$

We note, in passing, that a special case of our server-interruption model, where both the up-periods and the down-periods are geometrically distributed, could, in principle, be analyzed by means of the methodology developed in [59], because in that case the total number of available servers of the queueing system is a Markov chain on the set $\{1, 2\}$. Here, however, we aim at a more general treatment for which that methodology no longer works.

The fraction of time during which both servers are available, i.e., during which the

extra server is available as well, is given by

$$\begin{aligned}\sigma &\triangleq \frac{E[\text{up-period}]}{E[\text{up-period}] + E[\text{down-period}]} \\ &= \frac{\frac{1}{1-\alpha}}{\frac{1}{1-\alpha} + \bar{r}} = \frac{1}{1 + (1-\alpha)\bar{r}} \ .\end{aligned} \tag{6}$$

We note that σ can also be interpreted as the long-term probability that an arbitrary slot is an up-slot, whereas $1-\sigma$ corresponds to the probability that an arbitrary slot is a down-slot. We further emphasize that, in our model, the quantity σ can only take values *strictly smaller* than 1 (because $\bar{r} \geq 1$ and the parameter α, introduced in (3), is strictly smaller than 1) and *strictly greater* than zero. The limiting case $\sigma \to 1$ is obtained when $\bar{r}$ remains finite and $\alpha \to 1$ and corresponds to the situation where the extra server is (almost) always available (i.e., the mean down-period is negligible relative to the mean up-period), whereas the limiting case $\sigma \to 0$ is obtained when $\alpha < 1$ and $\bar{r} \to \infty$ and corresponds to the situation where the extra server is (almost) never available (i.e., the mean up-period is negligible relative to the mean down-period). In the sequel, we mainly deal with the "regular" case where $0 < \sigma < 1$; some attention is devoted to the limiting cases $\sigma \to 1$ and $\sigma \to 0$ in section 6.

An interesting alternative view to the above interruption model of the extra server is obtained if one concentrates on the relative position of two consecutive up-slots on the time axis, and the "discrete distance" (expressed in slots) between them, in the sequel referred to as the "inter-extra time". Consider an arbitrary ("tagged") up-slot on the time axis. Either this slot was the last slot of its up-period, which happens with probability $1-\alpha$, and then the time until the next up-slot on the time axis is equal to $1+\tilde{r}$, where $\tilde{r}$ refers to the length of the down-period that starts after the tagged up-slot. Or, else, with probability α, the tagged up-slot is not the last slot of its up-period, and then the time until the next up-slot on the time axis is simply 1 slot. It easily follows that the pgf of the time between two consecutive up-slots on the time axis, i.e., the pgf of the inter-extra time, is given by

$$F(z) \triangleq z[\alpha + (1-\alpha)R(z)] \ . \tag{7}$$

It is also clear that the consecutive inter-extra times are a set of i.i.d. random variables with common pgf $F(z)$, defined in (7). This condition also determines our server interruption model unambiguously and can be considered as an equivalent mathematical description of it. In the sequel we will use both descriptions according to convenience. Note that the mean value of the inter-extra time is given by

$$E[f] \triangleq F'(1) = 1 + (1-\alpha)R'(1) = 1 + (1-\alpha)\bar{r} = \frac{1}{\sigma} \ , \tag{8}$$

equivalent to $\sigma = 1/E[f]$, which is also intuitively clear. The second derivative of $F(z)$ at $z = 1$ is given by

$$F''(1) = \frac{2(1-\sigma)}{\sigma} + (1-\alpha)R''(1) \ . \tag{9}$$

In the sequel, it will turn out convenient to use a separate notation for the function

$$\hat{F}(z) \triangleq F\big(\frac{C(z)}{z}\big) \ . \tag{10}$$

It is easily seen that the first two derivatives of $\hat{F}(z)$ at $z = 1$ are given by

$$\hat{F}'(1) = \frac{\lambda - 1}{\sigma} \tag{11}$$

and

$$\hat{F}''(1) = \frac{C''(1) - 2(\lambda - 1)}{\sigma} + (\lambda - 1)^2 F''(1) \ . \tag{12}$$

There are at least three good reasons to opt for a geometric distribution for the up-periods. First, this choice apparently simplifies the analysis of our queueing model considerably, as will become clear in later sections. Second, as shown above, geometric up-periods imply i.i.d. inter-extra times. Third, if the extra server is, in fact, the regular server of a second similar queue, where arrivals occur according to an independent arrival process with pgf $\hat{C}(z)$, similar to the model in (1), then the idle times of the second queue – which are the up-periods of the extra server in the first queue – are, in fact, geometrically distributed, notably with parameter $\hat{C}(0)$. As for the down-periods, or, equivalently, the inter-extra times, we require no specific assumptions on the precise nature of the pgf $R(z)$, or $F(z)$, for the first - theoretical - part of the analysis, as will become clear further on, but the derivation of practical results, including the determination of a number of remaining unknown parameters in our formulas, turns out to be feasible only when further restrictions are imposed on $R(z)$. As far as we have been able to discover, the most general assumption for which the full analysis can be accomplished, seems to be that $R(z)$, or $F(z)$, must be *rational* functions of their argument.

The structure of the rest of this paper is as follows. In section 3, we analyze the steady-state queueing performance of the system, which results in exact expressions for the pgfs and the mean values of the numbers of customers in the system at various observation epochs. These formulas, however, still contain a theoretically infinite number of unknowns, which remain to be determined. In section 4, we discuss how these unknowns can be computed in case the pgf $R(z)$ of the down-periods is rational. Section 5 considers a number of specific choices for the arrival pgf $C(z)$, for which remarkable special instances of our model are obtained. Section 6 focuses on the cases $\sigma = 0$ and $\sigma = 1$ and also examines to what extent the corresponding results can be obtained by considering the limits $\sigma \to 0$ and $\sigma \to 1$ in our earlier findings. In section 7, we consider an extended special case, in order to investigate the impact of the precise details of the server-interruption model (i.e., mean availability σ, mean and coefficient of variation of the down-period distribution) on the performance of the system, and illustrate our results by means of some numerical examples. Section 8 states some conclusions and indicates some possible directions for future work.

3. Steady-state queueing analysis

3.1. Stability condition of the system

In the next subsections we will analyze the steady-state behavior of the queueing system under study. Before tackling this analysis, we first examine the conditions under which such a steady state exists. In general terms, it is not difficult to see that the system is stable, i.e., a steady state exists, if and only if the traffic intensity, i.e., the average

amount of *work* entering the system per slot, is strictly less than the average "service capacity" of the system, i.e., the average amount of *work* that the servers are able to deliver per slot when the system is saturated, i.e., when there are always customers available in the system. As in the system under study each customer brings in one unit of work, the traffic intensity is identical to the mean arrival rate λ, whereas the average service capacity of the system is simply given by $1+\sigma$, as the regular server can process one unit of work per time slot and the extra server can, on average, handle σ work units per slot. The stability condition thus reads

$$\lambda < 1 + \sigma \ , \tag{13}$$

where σ was defined in (6).

3.2. System evolution during up-periods and down-periods

Let g_0 (with pgf $G_0(z)$) and h_0 (with pgf $H_0(z)$) denote the total *system content*, i.e., the total number of customers present in the system (i.e., queue + servers) at the beginning of an up-period and a down-period, respectively, when the system has reached a steady state. Furthermore, for all $k \geq 1$, let g_k (with pgf $G_k(z)$) and h_k (with pgf $H_k(z)$), indicate the steady-state system content just after the kth slot (i.e., at the beginning of the $(k+1)$st slot) of an up-period and a down-period, respectively. The random variables g_k and h_k (for $k \geq 0$) are graphically illustrated in Fig. 1. Finally, let g (with pgf $G(z)$), h (with pgf $H(z)$) and u (with pgf $U(z)$) denote the steady-state system content just after an arbitrary up-slot, just after an arbitrary down-slot and just after a completely arbitrary time slot, respectively.

As during up-periods two servers are available, the following system equation is valid between the random variables $\{g_k\}$:

$$g_k = (g_{k-1} - 2)^+ + c_{\text{up},k} \ , \quad k \geq 1 \ , \tag{14}$$

where $(.)^+ \triangleq \max(.,0)$. Likewise, during down-periods only the regular server is available, and therefore the following system equation for the $\{h_k\}$s is valid:

$$h_k = (h_{k-1} - 1)^+ + c_{\text{down},k} \ , \quad k \geq 1 \ . \tag{15}$$

Here the quantities $c_{\text{up},k}$ and $c_{\text{down},k}$ refer to the number of arrivals during the kth slot of an up-period and a down-period, respectively.

Equations (14) and (15) can be translated into corresponding equations in the z-domain, by taking the pgfs of (14) and (15), which results in

$$G_k(z) \triangleq E[z^{g_k}] = \frac{C(z)}{z^2}[G_{k-1}(z) + G_{k-1}(0)(z^2-1) + G'_{k-1}(0)z(z-1)] \ , \quad k \geq 1 \ , \tag{16}$$

and

$$H_k(z) \triangleq E\big[z^{h_k}\big] = \frac{C(z)}{z}[H_{k-1}(z) + H_{k-1}(0)(z-1)] \ , \quad k \geq 1 \ , \tag{17}$$

respectively. Here the prime is used to indicate the first derivative.

Let us define the bivariate functions $G(x,z)$ and $H(x,z)$ as

$$G(x,z) \triangleq \sum_{k=0}^{\infty} x^k G_k(z) \quad ; \quad H(x,z) \triangleq \sum_{k=0}^{\infty} x^k H_k(z) \ . \tag{18}$$

Then, from (16), it follows that

$$G(x,z) = G_0(z) + \frac{C(z)}{z^2}\sum_{k=1}^{\infty} x^k[G_{k-1}(z) + G_{k-1}(0)(z^2-1) + G'_{k-1}(0)z(z-1)]$$

$$= G_0(z) + \frac{xC(z)}{z^2}[G(x,z) + (z^2-1)G(x,0) + z(z-1)\frac{\partial G}{\partial z}(x,0)] \ ,$$

from which $G(x,z)$ can be expressed as

$$G(x,z) = \frac{z^2 G_0(z) + x(z-1)C(z)[(z+1)G(x,0) + z\frac{\partial G}{\partial z}(x,0)]}{z^2 - xC(z)} \ . \tag{19}$$

Likewise, from (17), we can derive

$$H(x,z) = H_0(z) + \frac{C(z)}{z}\sum_{k=1}^{\infty} x^k[H_{k-1}(z) + H_{k-1}(0)(z-1)]$$

$$= H_0(z) + \frac{xC(z)}{z}[H(x,z) + (z-1)H(x,0)] \ ,$$

which results in

$$H(x,z) = \frac{zH_0(z) + x(z-1)C(z)H(x,0)}{z - xC(z)} \ . \tag{20}$$

Equation (20) can be inverted with respect to the variable x, which yields

$$H_k(z) = \left(\frac{C(z)}{z}\right)^k H_0(z) + (z-1)\sum_{i=1}^{k}\left(\frac{C(z)}{z}\right)^i H_{k-i}(0) \ , \quad k \geq 1 \ . \tag{21}$$

We note that the same result can also be easily obtained by recursive application of the original equation (17).

3.3. System content at the start of up-periods and down-periods: the pgfs $G_0(z)$ and $H_0(z)$

The next step in our analysis consists of establishing equations between the pgfs $G_0(z)$ and $H_0(z)$. We note that the random variable g_0 can be interpreted as the system content after the *last* slot of a down-period (see Fig. 1). In view of the definition of h_k and equation (5), we thus have, by the law of total probability

$$G_0(z) = \sum_{k=1}^{\infty} r(k)H_k(z) \ . \tag{22}$$

Using (21), this can be rewritten as

$$G_0(z) = H_0(z)\sum_{k=1}^{\infty} r(k)\Big(\frac{C(z)}{z}\Big)^k + (z-1)\sum_{k=1}^{\infty} r(k)\sum_{i=1}^{k}\Big(\frac{C(z)}{z}\Big)^i H_{k-i}(0)$$

$$= H_0(z)R\Big(\frac{C(z)}{z}\Big) + (z-1)\sum_{i=1}^{\infty}\Big(\frac{C(z)}{z}\Big)^i q(i)$$

$$= H_0(z)R\Big(\frac{C(z)}{z}\Big) + (z-1)Q\Big(\frac{C(z)}{z}\Big) \ , \tag{23}$$

where the unknown quantities $q(i), i \geq 1$, and their transform function $Q(y)$ are defined as

$$q(i) \triangleq \sum_{j=0}^{\infty} H_j(0)r(i+j) \quad ; \quad Q(y) \triangleq \sum_{i=1}^{\infty} q(i)y^i \ . \tag{24}$$

Similarly, we can view the random variable h_0 as the system content after the *last* slot of an up-period (see Fig. 1), which, according to equation (3), results in

$$H_0(z) = \sum_{k=1}^{\infty}(1-\alpha)\alpha^{k-1}G_k(z) \ . \tag{25}$$

Comparing equations (25) and (18), we easily conclude that

$$H_0(z) = \frac{1-\alpha}{\alpha}[G(\alpha,z) - G_0(z)] \ ,$$

or, using equation (19),

$$H_0(z) = \frac{(1-\alpha)C(z)}{z^2-\alpha C(z)}\big\{G_0(z) + (z-1)\big[(z+1)G(\alpha,0) + z\frac{\partial G}{\partial z}(\alpha,0)\big]\big\}$$

$$= \frac{C(z)}{z^2-\alpha C(z)}\{(1-\alpha)G_0(z) + (z-1)[p(0)+p(1)z]\} \ , \tag{26}$$

where the unknown quantities $p(0)$ and $p(1)$ are defined as

$$p(0) \triangleq (1-\alpha)G(\alpha,0) \quad ; \quad p(1) \triangleq (1-\alpha)[G(\alpha,0) + \frac{\partial G}{\partial z}(\alpha,0)] \ . \tag{27}$$

Equations (23) and (26) provide a set of two linear equations for the two pgfs $G_0(z)$ and $H_0(z)$, which can therefore be computed from this set. The resulting formulas are

$$H_0(z) = \frac{(z-1)C(z)[p(0)+p(1)z+(1-\alpha)Q\big(\frac{C(z)}{z}\big)]}{z^2 - C(z)[\alpha+(1-\alpha)R\big(\frac{C(z)}{z}\big)]} \tag{28}$$

and

$$G_0(z) = \frac{(z-1)\{[p(0)+p(1)z]C(z)R\big(\frac{C(z)}{z}\big) + [z^2-\alpha C(z)]Q\big(\frac{C(z)}{z}\big)\}}{z^2 - C(z)[\alpha+(1-\alpha)R\big(\frac{C(z)}{z}\big)]} \ , \tag{29}$$

in which only the unknown parameters $p(0)$ and $p(1)$, defined in (27), and $q(i), i \geq 1$, defined in (24), remain to be determined.

We note that the pgfs $G_0(z)$ and $H_0(z)$ can also be expressed in terms of the pgf $F(z)$ of the inter-extra times, or the related pgf $\hat{F}(z)$, defined in (7) and (10), as follows:

$$G_0(z) = \frac{(z-1)}{(1-\alpha)[z-\hat{F}(z)]}\left[\left[\hat{F}(z) - \alpha\frac{C(z)}{z}\right]\left[p(0)+p(1)z\right] + (1-\alpha)\left[z - \alpha\frac{C(z)}{z}\right]Q\left(\frac{C(z)}{z}\right)\right] \ , \tag{30}$$

$$H_0(z) = \frac{(z-1)}{z-\hat{F}(z)}\left(\frac{C(z)}{z}\right)\left[p(0)+p(1)z + (1-\alpha)Q\left(\frac{C(z)}{z}\right)\right] \ , \tag{31}$$

which will turn out to be useful expressions later on.

3.4. System content just after up-slots and down-slots: the pgfs $G(z)$ and $H(z)$

Let $S_{\rm up}$ and $S_{\rm down}$ indicate an arbitrary up-slot and an arbitrary down-slot respectively. Then the random variables g and h, defined in subsection 3.2, can be viewed as the steady-state system content just after $S_{\rm up}$ and $S_{\rm down}$, respectively. Let the random variables $K_{\rm up}$ and $K_{\rm down}$ indicate the ordinal numbers of slots $S_{\rm up}$ and $S_{\rm down}$ within the up-period and the down-period they belong to, respectively, i.e., $S_{\rm up}$ is the $K_{\rm up}$th slot of an up-period and $S_{\rm down}$ is the $K_{\rm down}$th slot of a down-period. It is well-known (see, e.g., [64]) that $K_{\rm up}$ and $K_{\rm down}$ have the following pmfs:

$$\text{Prob}[K_{\rm up}=k] = \frac{\text{Prob}[\text{ up-period } \geq k]}{E[\text{up-period}]} = \frac{\sum_{n=k}^{\infty}(1-\alpha)\alpha^{n-1}}{(1/1-\alpha)} = (1-\alpha)\alpha^{k-1} \ , \tag{32}$$

i.e., the random variable $K_{\rm up}$ is geometrically distributed with parameter α, just as the up-periods themselves, a consequence of the memoryless property of the geometric distribution, and

$$\text{Prob}[K_{\rm down}=k] = \frac{\text{Prob}[\text{ down-period } \geq k]}{E[\text{down-period}]} = \frac{\sum_{n=k}^{\infty} r(n)}{R'(1)} \ , \tag{33}$$

with corresponding pgf

$$K_{\rm down}(z) \triangleq E\left[z^{K_{\rm down}}\right] = \frac{z[R(z)-1]}{(z-1)R'(1)} \ . \tag{34}$$

From the definitions of $S_{\rm up}$ and $S_{\rm down}$, on the one hand, and the random variables $K_{\rm up}$ and $K_{\rm down}$, on the other hand, it is clear now that the random variables g and h can be viewed as the steady-state system contents just after the $K_{\rm up}$th slot of an up-period and just after the $K_{\rm down}$th slot of a down-period, respectively. Their pgfs $G(z)$ and $H(z)$ can therefore be computed in a very similar manner as the pgfs $H_0(z)$ and $G_0(z)$, respectively, in the previous subsection. Specifically, the pgf $G(z)$ is given by

$$G(z) = \sum_{k=1}^{\infty}\text{Prob}[K_{\rm up}=k]\,G_k(z) = \sum_{k=1}^{\infty}(1-\alpha)\alpha^{k-1}G_k(z) = H_0(z) \ , \tag{35}$$

in view of (32) and (25). The pgf $H(z)$, on the other hand, can be expressed as

$$H(z) = \sum_{k=1}^{\infty} \text{Prob}[K_{\text{down}} = k]\, H_k(z)$$

$$= H_0(z) K_{\text{down}}\left(\frac{C(z)}{z}\right) + (z-1) S\left(\frac{C(z)}{z}\right) \ , \tag{36}$$

which has to be compared to (22) and (23). Here the function $S(y)$ is defined in a similar way as the function $Q(y)$ in equation (24):

$$S(y) \triangleq \sum_{i=1}^{\infty} s(i) y^i \quad ; \quad s(i) \triangleq \sum_{j=0}^{\infty} H_j(0) \text{Prob}[K_{\text{down}} = i+j] \ . \tag{37}$$

Moreover, it turns out that the (unknown) function $S(y)$ can be expressed in terms of the (also unknown) function $Q(y)$, as follows:

$$S(y) = \sum_{i=1}^{\infty} y^i \sum_{j=0}^{\infty} H_j(0) \text{Prob}[K_{\text{down}} = i+j]$$

$$= \sum_{i=1}^{\infty} y^i \sum_{j=0}^{\infty} H_j(0) \sum_{n=i+j}^{\infty} \frac{r(n)}{R'(1)} = \sum_{i=1}^{\infty} y^i \sum_{j=0}^{\infty} H_j(0) \sum_{m=i}^{\infty} \frac{r(m+j)}{R'(1)}$$

$$= \frac{1}{R'(1)} \sum_{i=1}^{\infty} y^i \sum_{m=i}^{\infty} q(m) = \frac{1}{R'(1)} \sum_{m=1}^{\infty} q(m) \sum_{i=1}^{m} y^i$$

$$= \frac{1}{R'(1)} \sum_{m=1}^{\infty} q(m) \frac{y(y^m - 1)}{y-1} = \frac{y[Q(y) - Q(1)]}{(y-1)R'(1)} \ , \tag{38}$$

where we have used the expression for $\text{Prob}[K_{\text{down}} = k]$ in equation (33) and the quantities $q(m)$ and $Q(y)$ were defined in equation (24). Using (38) and (34) in (36), we then find the following expression for the pgf $H(z)$:

$$H(z) = \frac{C(z)}{R'(1)[C(z) - z]} \left\{ H_0(z) \left[R\left(\frac{C(z)}{z}\right) - 1 \right] + (z-1) \left[Q\left(\frac{C(z)}{z}\right) - Q(1) \right] \right\} \ . \tag{39}$$

3.5. System content at the start of an arbitrary slot: the pgf $U(z)$

In discrete-time queueing models, one is usually interested in the steady-state distribution of the system content at random slot boundaries, i.e., *at the beginning of* an arbitrary slot, or, equivalently, *just after* an arbitrary slot. Now, let S indicate such an arbitrary slot in steady state. Then with probability σ, S is an up-slot, whereas it is a down-slot with probability $1 - \sigma$; here the quantity σ was defined in equation (6). It then easily follows that the pgf $U(z)$ of the steady-state system content at random slot boundaries is given by

$$U(z) = \sigma G(z) + (1 - \sigma) H(z) \ , \tag{40}$$

where the pgfs $G(z)$ and $H(z)$ were determined in the previous subsection. From equations (35) and (39), it then follows that $U(z)$ can be expressed as

$$U(z) = \frac{z-1}{z-C(z)}\big\{[1+\sigma-\lambda+\sigma p(1)(z-1)]C(z) - \sigma z H_0(z)\big\} \ , \tag{41}$$

or, using equation (28) for the pgf $H_0(z)$,

$$U(z) = \frac{(z-1)C(z)}{z-C(z)}\left\{1+\sigma-\lambda-\sigma(z-1)\frac{p(0)z + p(1)C(z)[\alpha+(1-\alpha)R\big(\frac{C(z)}{z}\big)] + (1-\alpha)zQ\big(\frac{C(z)}{z}\big)}{z^2 - C(z)[\alpha+(1-\alpha)R\big(\frac{C(z)}{z}\big)]}\right\} \ . \tag{42}$$

We note, again, that equation (42) completely expresses the pgf $U(z)$ in terms of the original model parameters, on the one hand, and the unknown parameters $p(0)$ and $p(1)$, defined in (27), and $q(i), i \geq 1$, defined in (24), on the other hand. In terms of the pgf $\hat{F}(z)$ defined in (10), $U(z)$ can also be very nicely expressed as

$$U(z) = \frac{(z-1)C(z)}{z-C(z)}\left\{(1+\sigma-\lambda) - \frac{\sigma(z-1)}{z-\hat{F}(z)}\left[p(0)+p(1)\hat{F}(z)+(1-\alpha)Q\left(\frac{C(z)}{z}\right)\right]\right\} \ . \tag{43}$$

3.6. Mean system content at various observation epochs

From a practical point of view, the most important results obtained so far, are the pgfs $G_0(z)$, $H_0(z)$ and $U(z)$, because they describe the system-content distributions at observation epochs of practical interest, i.e., at the beginning of an up-period (or, equivalently, just after a down-period, when accumulation of customers is more likely), at the beginning of a down-period (i.e., just after an up-period, when the service capacity has been high for a while), and at random slot boundaries (which gives an overall view), respectively. In particular, the corresponding mean values have practical relevance. These can be obtained by applying the moment-generating property of pgfs on equations (30), (31) and (43). The results are summarized below.

$$E[h_0] = H_0'(1) = \lambda - 1 + \frac{\sigma}{1+\sigma-\lambda}\Big[p(1) + (1-\alpha)(\lambda-1)Q'(1) + \hat{F}''(1)\Big] \ ;$$

$$E[g_0] = G_0'(1) = (\lambda-1)\frac{1-\alpha\sigma}{\sigma(1-\alpha)} + Q(1) + \frac{\sigma}{1+\sigma-\lambda}\Big[p(1) + (1-\alpha)(\lambda-1)Q'(1) + \hat{F}''(1)\Big] \ ;$$

$$E[u] = U'(1) = \lambda + \frac{C''(1)}{2(1-\lambda)} + \frac{\sigma^2\hat{F}''(1)}{2(\lambda-1)(1+\sigma-\lambda)} + \frac{\sigma}{1+\sigma-\lambda}[p(1) + (1-\alpha)\sigma Q'(1)] \ . \tag{44}$$

4. Determining the remaining unknowns

In section 3, we have analyzed the queueing system under consideration without making any specific assumptions as to the nature of the pgf $R(z)$ of the down-periods, or, equivalently, the pgf $F(z)$ of the inter-extra times. As a result, we have been able

to derive expressions for the pgfs and the mean values of the system content at various observation epochs, in terms of the original system parameters ($C(z)$, $R(z)$, α), or, equivalently, ($C(z)$, $F(z)$), and a number of unknown parameters, i.e., the quantities $p(0)$ and $p(1)$, defined in (27), and $q(i), i \geq 1$, defined in (24). The determination of these remaining unknowns is the objective of the current section.

As in most queueing analyses, a first equation for the remaining unknowns can be retrieved by expressing the normalization condition for the probability distribution of the system content. In terms of the pgfs that we obtained earlier, this condition can be expressed by requiring that the pgfs should return the value 1 when the argument z is replaced by the value 1. It is not difficult to see that the result does not depend on which specific pgf of the system content is selected. Choosing $H_0(z)$, as given by equation (28), we obtain

$$\lim_{z \to 1} H_0(z) = \frac{p(0) + p(1) + (1-\alpha)Q(1)}{2 - \lambda - (1-\alpha)R'(1)(\lambda - 1)} = 1 \ ,$$

from which it easily follows that

$$p(0) + p(1) + (1-\alpha)Q(1) = \frac{1 + \sigma - \lambda}{\sigma} \ , \tag{45}$$

where we have introduced the parameter σ according to (6).

At first sight, it seems as if the number of remaining unknowns is infinitely large, as the quantities $q(i)$ appear in our results for all $i \geq 1$. However, it turns out that not all of the parameters $q(i), i \geq 1$, are necessarily linearly independent. Specifically, we show in the rest of this section that only a finite number of independent unknowns remains in case the pgf $R(z)$ is a rational function of its argument z.

4.1. Notations

From now on, let us assume the pgf $R(z)$ is a rational function of z, which can be expressed as the ratio of two normalized and mutually prime polynomials $A(z)$ and $B(z)$, i.e.,

$$R(z) = \frac{A(z)}{B(z)} \ , \tag{46}$$

with

$$A(z) = \sum_{i=1}^{m_A} a_i z^i \quad ; \quad B(z) = \sum_{j=0}^{m_B} b_j z^j \ , \tag{47}$$

where

$$A(1) = \sum_{i=1}^{m_A} a_i = 1 \tag{48}$$

and

$$B(1) = \sum_{j=0}^{m_B} b_j = 1 \ . \tag{49}$$

Note that, in view of the fact that $R(z)$ is a pgf, the polynomial $B(z)$ has no zeroes inside the closed unit disk $\{z : |z| \leq 1\}$ and exactly m_B zeroes outside the closed unit disk in the complex z-plane. Moreover, as each down-period lasts at least one slot, we also have

$$R(0) = A(0) = 0 \ . \tag{50}$$

Let us also define the parameter $\hat{m}$ as

$$\hat{m} \triangleq \max\{m_A, m_B\} \ . \tag{51}$$

4.2. The unknown function $Q(y)$

The unknown parameters $q(i)$ and their transform $Q(y)$ can be computed from equation (24) as

$$q(i) \triangleq \sum_{j=0}^{\infty} H_j(0) r(i+j) \tag{52}$$

and

$$Q(y) \triangleq \sum_{i=1}^{\infty} q(i) y^i = \sum_{i=1}^{\infty} y^i \sum_{j=0}^{\infty} H_j(0) r(i+j) = \sum_{j=0}^{\infty} H_j(0) R_j(y) \ , \tag{53}$$

where the functions $R_j(y), j \geq 0$, are defined as

$$R_j(y) \triangleq \sum_{i=1}^{\infty} r(i+j) y^i \ . \tag{54}$$

We will now show that all the functions $R_j(y)$ are rational functions of their argument y with the same denominator $B(y)$ as the pgf $R(y)$, implying that the same holds for the function $Q(y)$, which, according to (53), is simply a linear combination of the functions $R_j(y)$. In order to prove this property, we first define the functions $L_j(y), j \geq 0$, as

$$L_j(y) \triangleq B(y) R_j(y) \ , \ \ j \geq 0 \ . \tag{55}$$

If we can prove that the functions $L_j(y)$ are polynomials for all values of $j \geq 0$, then we are done. We will do even a little more than this and prove that all the functions $L_j(y), j \geq 0$, are polynomials of maximum degree $\hat{m}$, as defined in (51), divisible by a factor y, i.e., for which $L_j(0) = 0$. We construct a proof by induction on j.

The basis step of the proof consists of observing that the statement is true for $j = 0$. From the above definitions, it is immediately clear that

$$R_0(y) = R(y) \ , \tag{56}$$

and, therefore, in view of (46),

$$L_0(y) \triangleq B(y) R(y) = A(y) \ , \tag{57}$$

which indeed is a polynomial of maximum degree $\hat{m}$, divisible by y, i.e., for which $L_0(0) = 0$.

In order to prove the induction step, we assume that the statement is true for a value j, i.e., we assume that the function $L_j(y)$ is a polynomial of maximum degree $\hat{m}$, divisible by y. Based on this assumption, we then show that the same holds for the

function $L_{j+1}(y)$. From equations (54) and (55), it follows that

$$\begin{aligned} L_{j+1}(y) \triangleq B(y)R_{j+1}(y) &= B(y)\sum_{i=1}^{\infty} r(i+j+1)y^i \\ &= B(y)\left[\sum_{i=0}^{\infty} r(i+j+1)y^i - r(j+1)\right] \\ &= B(y)\left[\frac{R_j(y)}{y} - r(j+1)\right] \\ &= \frac{L_j(y)}{y} - r(j+1)B(y) \ . \end{aligned} \tag{58}$$

By assumption, $L_j(y)$ is a polynomial of maximum degree $\hat{m}$, divisible by y, whereas $B(y)$ is a polynomial of degree m_B and, hence, also of maximum degree $\hat{m}$. Therefore, the above equation shows that $L_{j+1}(y)$ is also a polynomial of maximum degree $\hat{m}$. Moreover, $L_{j+1}(y)$ is divisible by the factor y, as follows from

$$L_{j+1}(0) = B(0)\left[\lim_{y\to 0}\frac{R_j(y)}{y} - r(j+1)\right] = B(0)[r(j+1) - r(j+1)] = 0 \ . \tag{59}$$

From equation (53) it now follows that

$$B(y)Q(y) = \sum_{j=0}^{\infty} H_j(0)B(y)R_j(y) = \sum_{j=0}^{\infty} H_j(0)L_j(y) \triangleq \sum_{i=1}^{\hat{m}} \hat{a}_i y^i \ , \tag{60}$$

in view of the fact that all the functions $L_j(y), j \geq 0$, are polynomials of maximum degree $\hat{m}$, divisible by a factor y. As a result, the unknown function $Q(y)$ can be expressed as

$$Q(y) = \frac{\hat{A}(y)}{B(y)} \ , \tag{61}$$

with

$$\hat{A}(y) \triangleq \sum_{i=1}^{\hat{m}} \hat{a}_i y^i \quad ; \quad B(y) = \sum_{j=0}^{m_B} b_j y^j \ . \tag{62}$$

Equations (61) and (62) show that the function $Q(y)$ contains only $\hat{m}$ independent unknowns, i.e., the parameters $\hat{a}_i, 1 \leq i \leq \hat{m}$. The formal resemblance between the expressions (46) for $R(z)$ and (61) for $Q(y)$ is striking and also turns out to be the key to the determination of the remaining unknowns in case the pgf $R(z)$ is rational. We note, however, that, as opposed to $R(z)$, the function $Q(y)$ is not necessarily normalized; specifically, (61) and (49) imply that

$$Q(1) = \hat{A}(1) = \sum_{i=1}^{\hat{m}} \hat{a}_i \ . \tag{63}$$

4.3. Finding equations for the remaining unknowns

Using the above results in equation (28), we now find the following expression for the pgf $H_0(z)$:

$$H_0(z) = \frac{(z-1)C(z)\big\{[p(0)+p(1)z]B\big(\frac{C(z)}{z}\big) + (1-\alpha)\hat{A}\big(\frac{C(z)}{z}\big)\big\}}{[z^2 - \alpha C(z)]B\big(\frac{C(z)}{z}\big) - (1-\alpha)C(z)A\big(\frac{C(z)}{z}\big)} \quad . \tag{64}$$

The above expression is dependent on the variable z in two distinct ways: either through the combination $C(z)/z$, either through z itself. For the sake of simplifying the computations, it turns out to be useful to introduce a separate notation for the combination $C(z)/z$, as follows:

$$x \triangleq \left(\frac{C(z)}{z}\right)^{-1} = \frac{z}{C(z)} \quad . \tag{65}$$

Equation (65) basically defines x for any value of z. On the other hand, for any given value of x within the unit disk $\{x : |x| \leq 1\}$ of the complex x-plane, there is also exactly one value of z within the unit disk $\{z : |z| \leq 1\}$ of the complex z-plane that satisfies equation (65), as can be readily shown by means of an application of Rouché's theorem from complex analysis [65, 64] on the function $z - xC(z)$. Let us indicate this specific value of z as $e(x)$. Then, clearly,

$$\begin{aligned} &e(x) = xC(e(x)) \quad , \quad \text{for all } x \in \{x : |x| \leq 1\} \ ; \\ &e\left(\frac{z}{C(z)}\right) = z \quad , \quad \text{for all } z \in \{z : |z| \leq 1\} \ ; \\ &e(0) = 0 \ ; \\ &e(1) = 1 \ . \end{aligned} \tag{66}$$

Choosing $z = e(x)$ in equation (64), we then obtain

$$H_0\big(e(x)\big) = \frac{[e(x)-1]\big\{[p(0)+p(1)e(x)]B(x^{-1}) + (1-\alpha)\hat{A}(x^{-1})\big\}}{[xe(x)-\alpha]B(x^{-1}) - (1-\alpha)A(x^{-1})} \quad . \tag{67}$$

Here (47) and (62) imply that the functions $A(x^{-1})$, $B(x^{-1})$ and $\hat{A}(x^{-1})$ can be expressed in terms of nonnegative powers of x as follows:

$$A(x^{-1}) = \frac{F_A(x)}{x^{m_A}} \quad ; \quad B(x^{-1}) = \frac{F_B(x)}{x^{m_B}} \quad ; \quad \hat{A}(x^{-1}) = \frac{F_{\hat{A}}(x)}{x^{\hat{m}}} \quad , \tag{68}$$

where the polynomials $F_A(x)$, $F_B(x)$ and $F_{\hat{A}}(x)$ are defined as

$$F_A(x) \triangleq \sum_{i=1}^{m_A} a_i x^{m_A - i} \quad ; \quad F_B(x) \triangleq \sum_{j=0}^{m_B} b_j x^{m_B - j} \quad ; \quad F_{\hat{A}}(x) \triangleq \sum_{i=1}^{\hat{m}} \hat{a}_i x^{\hat{m} - i} \quad . \tag{69}$$

Note that $F_A(x)$ and $F_B(x)$ are known polynomials of x, whereas $F_{\hat{A}}(x)$ is an unknown polynomial of x; furthermore $F_A(x)$ and $F_B(x)$ are normalized, i.e.,

$$F_A(1) = \sum_{i=1}^{m_A} a_i = 1 \quad ; \quad F_B(1) = \sum_{j=0}^{m_B} b_j = 1 \ , \tag{70}$$

in view of (48) and (49), while $F_{\hat{A}}(x)$ is not necessarily normalized:

$$F_{\hat{A}}(1) = \sum_{i=1}^{\hat{m}} \hat{a}_i = Q(1) \ , \tag{71}$$

in view of (63). Introducing new notations for the remaining unknowns related to $Q(y)$, i.e.,

$$\hat{p}(n) \triangleq (1-\alpha)\hat{a}_{\hat{m}-n} \quad , \quad 0 \le n \le \hat{m}-1 \quad ; \quad \hat{P}(x) \triangleq \sum_{n=0}^{\hat{m}-1} \hat{p}(n)x^n = (1-\alpha)F_{\hat{A}}(x) \ , \tag{72}$$

equation (67) can then be rewritten as

$$H_0\big(e(x)\big) = \frac{[e(x)-1]N(x)}{D(x)} \ , \tag{73}$$

where the (numerator) function $N(x)$ is given by

$$N(x) \triangleq \hat{P}(x) + [p(0) + p(1)e(x)]x^{\hat{m}-m_B}F_B(x) \tag{74}$$

and the denominator function $D(x)$ is defined as

$$D(x) \triangleq e(x)x^{\hat{m}+1-m_B}F_B(x) - \alpha x^{\hat{m}-m_B}F_B(x) - (1-\alpha)x^{\hat{m}-m_A}F_A(x) \ . \tag{75}$$

We note that the numerator $N(x)$ contains a finite number of remaining unknowns, i.e., the $\hat{m}+2$ quantities $p(0)$, $p(1)$, $\hat{p}(0), \hat{p}(1), \cdots, \hat{p}(\hat{m}-1)$. On the other hand, the denominator $D(x)$ of (73) is a known function of x, which has exactly $\hat{m}+2$ zeroes inside the closed unit disk $\{x : |x| \le 1\}$ of the complex x-plane. The proof of this property can be established by means of Rouché's theorem from complex analysis [65, 64], keeping in mind that the function $e(x)$, by its definition (see equation (66)), has exactly one zero inside the closed unit disk, i.e., $x = 0$, the factor $x^{\hat{m}+1-m_B}$ has an $(\hat{m}+1-m_B)$-fold zero at $x = 0$, and the polynomial $F_B(x)$, being defined in (68) as $F_B(x) \triangleq x^{m_B}B(x^{-1})$ has all its m_B zeroes inside the complex unit disk, since the polynomial $B(z)$ has all its zeroes outside the closed unit disk. This implies that the first term in the right hand side of equation (75) has $1 + (\hat{m}+1-m_B) + m_B = \hat{m}+2$ zeroes inside the closed unit disk; by Rouché's theorem, the same property holds for the whole right hand side of (75), i.e., for the function $D(x)$. It is not difficult to see that one of these zeroes is equal to $x = 1$, as

$$D(1) = e(1)\cdot 1 \cdot F_B(1) - \alpha \cdot 1 \cdot F_B(1) - (1-\alpha)\cdot 1 \cdot F_A(1) = 1 - \alpha - (1-\alpha) = 0 \ , \tag{76}$$

in view of (66) and (70). Now, for all values of x inside the closed unit disk in the complex x-plane, the quantity $e(x)$ is also inside the closed unit disk (by its very definition),

and therefore the function $H_0\big(e(x)\big)$ is bounded, which implies that the zeroes of its denominator $D(x)$ inside the closed unit disk must also be zeroes of its numerator $[e(x)-1]N(x)$. This property, combined with the normalization condition (45), in this case given by

$$p(0) + p(1) + \sum_{n=0}^{\hat{m}-1} \hat{p}(n) = \frac{1+\sigma-\lambda}{\sigma} \ , \tag{77}$$

yields a system of linear equations (in this case $\hat{m}+2$ linear equations for $\hat{m}+2$ unknowns) which can therefore – in general – be solved for the remaining unknowns $p(0)$, $p(1)$, $\hat{p}(0), \hat{p}(1), \cdots, \hat{p}(\hat{m}-1)$. In the next subsection, we show that explicit solution of this system of equations is not even necessary; it turns out that all results can, actually, be expressed directly in terms of the zeroes of $D(x)$ inside the closed unit disk.

4.4. Practical determination of the remaining unknowns

4.4.1. A useful polynomial

Let us indicate the $\hat{m}+2$ zeroes of $D(x)$ inside the closed unit disk $\{x : |x| \leq 1\}$ of the complex x-plane as $\{x_i, 0 \leq i \leq \hat{m}+1\}$, where, by convention, $x_0 = 1$. For $1 \leq i \leq \hat{m}+1$, we then have $D(x_i) = 0$ and $N(x_i) = 0$, and, hence, also

$$x_i N(x_i) - p(1) D(x_i) \triangleq V(x_i) = 0 \ , \tag{78}$$

where the function $V(x)$ is defined as

$$\begin{aligned} V(x) &\triangleq xN(x) - p(1)D(x) \\ &= x\hat{P}(x) + p(0)x^{\hat{m}+1-m_B}F_B(x) + p(1)\big[\alpha x^{\hat{m}-m_B}F_B(x) + (1-\alpha)x^{\hat{m}-m_A}F_A(x)\big] \ . \end{aligned} \tag{79}$$

As opposed to the functions $D(x)$ and $N(x)$, the newly defined function $V(x)$ is a polynomial function; in fact, $V(x)$ was constructed in such a way, so as to eliminate the (non-polynomial) quantity $e(x)$ from the expressions (74) and (75). Moreover, $V(x)$ is a polynomial of degree $\hat{m}+1$, the highest degree appearing in the second term of (79). Since a polynomial has exactly as many zeroes in the complex plane as its degree, and we know that all the quantities $\{x_i, 1 \leq i \leq \hat{m}+1\}$ are zeroes of $V(x)$, we can express $V(x)$ in terms of these zeroes in product form, as follows:

$$V(x) = \hat{V} \prod_{i=1}^{\hat{m}+1} (x - x_i) \ , \tag{80}$$

where $\hat{V}$ is a proportionality constant which remains to be determined, but the remaining part of $V(x)$ is fully known, as soon as the zeroes $\{x_i, 1 \leq i \leq \hat{m}+1\}$ have been computed. The value of $\hat{V}$ can be easily found by observing that, according to (80), (79), (74) and (77),

$$V(1) = \hat{V} \prod_{i=1}^{\hat{m}+1} (1 - x_i) = N(1) = p(0) + p(1) + \sum_{n=0}^{\hat{m}-1} \hat{p}(n) = \frac{1+\sigma-\lambda}{\sigma} \ , \tag{81}$$

and, hence,

$$\hat{V} = \frac{1+\sigma-\lambda}{\sigma \prod_{i=1}^{\hat{m}+1}(1-x_i)} \ , \tag{82}$$

so that $V(x)$ is given explicitly by

$$V(x) = \frac{1+\sigma-\lambda}{\sigma} \prod_{i=1}^{\hat{m}+1} \frac{x-x_i}{1-x_i} \ . \tag{83}$$

Having determined the above explicit expression for the polynomial $V(x)$, we are now in a good position to compute the remaining unknowns $p(0)$, $p(1)$ and $Q(y)$.

4.4.2. Determining $p(0)$

Expression (80) makes clear that $\hat{V}$ must be equal to the coefficient of the highest-degree term of $V(x)$, which, according to (79) is given by

$$\hat{V} = p(0)b_0 = p(0)B(0) \ , \tag{84}$$

where b_0 and $B(0)$ were defined in (47). It follows that the unknown parameter $p(0)$ can be expressed as

$$p(0) = \frac{\hat{V}}{b_0} = \frac{1+\sigma-\lambda}{\sigma b_0 \prod_{i=1}^{\hat{m}+1}(1-x_i)} \ . \tag{85}$$

4.4.3. Determining $p(1)$

Similarly, the unknown parameter $p(1)$ can be derived by computing $V(0)$ both from equation (79) and equation (80):

$$V(0) = -p(1)D(0) = \hat{V} \prod_{i=1}^{\hat{m}+1}(-x_i) \ , \tag{86}$$

so that

$$p(1) = -\frac{\hat{V}\prod_{i=1}^{\hat{m}+1}(-x_i)}{D(0)} = -\frac{1+\sigma-\lambda}{\sigma D(0)} \prod_{i=1}^{\hat{m}+1} \frac{-x_i}{1-x_i} \ . \tag{87}$$

Here $D(0)$ is a known quantity; its precise value depends on the relative values of the degrees m_A and m_B of the polynomials $A(z)$ and $B(z)$ in the definition (47) of the pgf $R(z)$. Specifically, putting $x = 0$ in (75) and keeping in mind the definition of $\hat{m}$ in (51), we have

$$\begin{aligned}
&D(0) = -\alpha F_B(0) = -\alpha b_{m_B} \quad , \quad \text{if } m_A < m_B \ ; \\
&D(0) = -\alpha F_B(0) - (1-\alpha)F_A(0) = -\alpha b_{m_B} - (1-\alpha)a_{m_A} \quad , \quad \text{if } m_A = m_B; \\
&D(0) = -(1-\alpha)F_A(0) = -(1-\alpha)a_{m_A} \quad , \quad \text{if } m_A > m_B \ .
\end{aligned} \tag{88}$$

Here a_{m_A} and b_{m_B} represent the coefficients of the highest-degree terms in the polynomials $A(z)$ and $B(z)$, respectively (see equations (47)).

4.4.4. Determining $Q(y)$

In order to compute the unknown function $Q(y)$, appearing in many of the formulas derived earlier, we first rewrite equation (79) as

$$V(x) = x\hat{P}(x) + p(0)x^{\hat{m}+1}B(x^{-1}) + p(1)x^{\hat{m}}\big[\alpha B(x^{-1}) + (1-\alpha)A(x^{-1})\big] \ , \tag{89}$$

where we have used (68) to reintroduce the polynomials $A(z)$ and $B(z)$. Keeping in mind the definition of these polynomials in (46) and the definition of the pgf $F(z)$ of the inter-extra times in (7), we can further express this as

$$V(x) = x\hat{P}(x) + x^{\hat{m}+1}B(x^{-1})[p(0) + p(1)F(x^{-1})] \ . \tag{90}$$

Moreover, equations (72), (68) and (61) imply that the functions $\hat{P}$ and Q are related as

$$\hat{P}(x) = (1-\alpha)F_{\hat{A}}(x) = (1-\alpha)x^{\hat{m}}\hat{A}(x^{-1}) = (1-\alpha)x^{\hat{m}}B(x^{-1})Q(x^{-1}) \ . \tag{91}$$

Combining (90) and (91), we then obtain

$$V(x) = x^{\hat{m}+1}B(x^{-1})\big[p(0) + p(1)F(x^{-1}) + (1-\alpha)Q(x^{-1})\big] \ . \tag{92}$$

Choosing $x = y^{-1}$ in the above result, we get

$$p(0) + p(1)F(y) + (1-\alpha)Q(y) = \frac{y^{\hat{m}+1}V(y^{-1})}{B(y)} = \frac{1+\sigma-\lambda}{\sigma B(y)}\prod_{i=1}^{\hat{m}+1}\frac{1-x_i y}{1-x_i} \ , \tag{93}$$

where we have used the known expression (83) for the polynomial $V(x)$. We note, in passing, that for $y = 1$, the above equation simply reduces to the normalization condition (45). From this result and the known expressions (85) and (87) for the parameters $p(0)$ and $p(1)$, the original unknown function $Q(y)$ can finally be expressed as

$$Q(y) = \frac{1+\sigma-\lambda}{(1-\alpha)\sigma\prod_{i=1}^{\hat{m}+1}(1-x_i)}\left[\frac{\prod_{i=1}^{\hat{m}+1}(1-x_i y)}{B(y)} - \frac{1}{b_0} + \frac{F(y)\prod_{i=1}^{\hat{m}+1}(-x_i)}{D(0)}\right] \ , \tag{94}$$

in which all quantities are known. The quantities $Q(1)$ and $Q'(1)$ appearing in some of our earlier results can be computed from (94) as

$$Q(1) = \frac{1+\sigma-\lambda}{(1-\alpha)\sigma}\left[1 - \frac{1}{\prod_{i=1}^{\hat{m}+1}(1-x_i)}\left(\frac{1}{b_0} - \frac{\prod_{i=1}^{\hat{m}+1}(-x_i)}{D(0)}\right)\right] \tag{95}$$

and

$$Q'(1) = \frac{1+\sigma-\lambda}{(1-\alpha)\sigma}\left[\sum_{i=1}^{\hat{m}+1}\frac{-x_i}{1-x_i} - B'(1) + \frac{1}{\sigma D(0)}\prod_{i=1}^{\hat{m}+1}\frac{-x_i}{1-x_i}\right] \ . \tag{96}$$

4.5. Final results

Having determined the remaining unknowns $p(0)$, $p(1)$ and $Q(y)$ in the previous subsection, we are now in a position to derive explicit closed-form expressions for all quantities of interest. Let us first consider the pgfs of the system content at various

observation epochs. Using (85), (87) and (94) (with $y = C(z)/z$) in equations (30) and (31), we get

$$G_0(z) = \frac{(1+\sigma-\lambda)(z-1)}{(1-\alpha)\sigma\prod_{i=1}^{\hat{m}+1}(1-x_i)}\left[\frac{\alpha C(z)}{D(0)z}\prod_{i=1}^{\hat{m}+1}(-x_i) - \frac{1}{b_0} + \frac{z^2-\alpha C(z)}{z[z-\hat{F}(z)]B\left(\frac{C(z)}{z}\right)}\prod_{i=1}^{\hat{m}+1}\frac{z-x_iC(z)}{z}\right] ; \tag{97}$$

$$H_0(z) = \frac{(1+\sigma-\lambda)(z-1)C(z)}{\sigma z\prod_{i=1}^{\hat{m}+1}(1-x_i)}\left[\frac{-1}{D(0)}\prod_{i=1}^{\hat{m}+1}(-x_i) + \frac{1}{[z-\hat{F}(z)]B\left(\frac{C(z)}{z}\right)}\prod_{i=1}^{\hat{m}+1}\frac{z-x_iC(z)}{z}\right] . \tag{98}$$

Remarkably, an even simpler expression can be derived for the pgf $U(z)$ of the system content at random slot boundaries, by plugging the expression (93) (with $y = C(z)/z$) directly in our earlier result (43):

$$U(z) = \frac{(1+\sigma-\lambda)(z-1)C(z)}{z-C(z)}\left[1 - \frac{z-1}{[z-\hat{F}(z)]B\left(\frac{C(z)}{z}\right)}\prod_{i=1}^{\hat{m}+1}\frac{z-x_iC(z)}{(1-x_i)z}\right] . \tag{99}$$

The corresponding expected values of the system content can be either derived directly from the above expressions for the pgfs, by computing the first derivatives at $z = 1$, or, alternatively, from our earlier expressions (44), by using equations (87), (95) and (96) for the quantities $p(1)$, $Q(1)$ and $Q'(1)$ respectively. The final formulas are

$$\begin{aligned} E[g_0] = \quad & \frac{C''(1)+\sigma(\lambda-1)^2F''(1)}{2(1+\sigma-\lambda)} + (\lambda-1)\left[\frac{\sigma-\lambda}{1+\sigma-\lambda} - \sum_{i=1}^{\hat{m}+1}\frac{x_i}{1-x_i} - B'(1)\right] \\ & + \frac{2-\lambda}{1-\alpha} + \frac{1+\sigma-\lambda}{(1-\alpha)\sigma\prod_{i=1}^{\hat{m}+1}(1-x_i)}\left[\frac{\alpha}{D(0)}\prod_{i=1}^{\hat{m}+1}(-x_i) - \frac{1}{b_0}\right] ; \end{aligned} \tag{100}$$

$$\begin{aligned} E[h_0] = \quad & \frac{C''(1)+\sigma(\lambda-1)^2F''(1)}{2(1+\sigma-\lambda)} + (\lambda-1)\left[\frac{\sigma-\lambda}{1+\sigma-\lambda} - \sum_{i=1}^{\hat{m}+1}\frac{x_i}{1-x_i} - B'(1)\right] \\ & - \frac{1+\sigma-\lambda}{\sigma D(0)}\prod_{i=1}^{\hat{m}+1}\frac{-x_i}{1-x_i} ; \end{aligned} \tag{101}$$

$$E[u] = \frac{C''(1)+\sigma^2(\lambda-1)F''(1)}{2(1+\sigma-\lambda)} + \frac{(\lambda-1)(\sigma-\lambda)}{1+\sigma-\lambda} - \sigma\left[\sum_{i=1}^{\hat{m}+1}\frac{x_i}{1-x_i} + B'(1)\right] . \tag{102}$$

4.6. Computing the zeroes $\{x_i\}$

The results in the previous subsection are expressed in terms of the original system parameters ($C(z)$, $R(z)$, α), or, equivalently, ($C(z)$, $F(z)$), on the one hand, and the

$\hat{m}+1$ zeroes of the function $D(x)$, defined in (75), strictly inside the (open) unit disk $\{x : |x| < 1\}$ of the complex x-plane, i.e., the quantities $\{x_i, 1 \leq i \leq \hat{m}+1\}$. Remember that, by convention $x_0 = 1$. In order to produce numerical results for various performance parameters of interest, it is thus necessary to solve the equation $D(x) = 0$ inside the open unit disk of the complex x-plane. This may seem a difficult task at first sight in view of the fact that $D(x)$ is given by

$$D(x) \triangleq e(x)x^{\hat{m}+1-m_B}F_B(x) - \alpha x^{\hat{m}-m_B}F_B(x) - (1-\alpha)x^{\hat{m}-m_A}F_A(x) \ ,$$

an expression which contains the function $e(x)$, defined only implicitly in (66) as

$$e(x) = xC(e(x)) \ , \quad \text{for all } x \in \{x : |x| \leq 1\} \ ;$$

$$e\Big(\frac{z}{C(z)}\Big) = z \ , \quad \text{for all } z \in \{z : |z| \leq 1\} \ .$$

In order to avoid this difficulty, it is useful to perform a change of variable in the equation $D(x) = 0$ by using the (bijective) transform (65) between the unit disks of the x-plane and the z-plane:

$$x = \frac{z}{C(z)} \quad \Longleftrightarrow \quad z = e(x) \ . \tag{103}$$

It is easily seen that, based on (68), (46), (7) and (10),

$$\begin{aligned} D(x) = 0 \quad &\Longleftrightarrow \quad xe(x)B(x^{-1}) = \alpha B(x^{-1}) + (1-\alpha)A(x^{-1}) \\ &\Longleftrightarrow \quad e(x) = x^{-1}\big[\alpha + (1-\alpha)R(x^{-1})\big] = F(x^{-1}) \\ &\Longleftrightarrow \quad z = F\big(\frac{C(z)}{z}\big) = \hat{F}(z) \ . \end{aligned}$$

It follows that the zeroes $\{x_i, 1 \leq i \leq \hat{m}+1\}$ can be expressed as

$$x_i = \frac{z_i}{C(z_i)} \ , \tag{104}$$

where the quantities $\{z_i, 1 \leq i \leq \hat{m}+1\}$ are the $\hat{m}+1$ roots of the equation $z = \hat{F}(z)$, strictly inside the (open) unit disk $\{z : |z| < 1\}$ of the complex z-plane. The latter equation does not contain any implicit functions anymore and may, therefore, be easier to solve than the original equation $D(x) = 0$.

5. Specific arrival distributions

So far, in this paper, we have made no specific assumptions or restrictions on the precise nature of the distribution of the number of arrivals per slot, i.e., our results are valid for all possible choices of the arrival pgf $C(z)$. In this section, we examine a number of special choices for the pgf $C(z)$, for which the system reduces to a remarkable special case, which is either interesting in its own right or which can serve as a check on the correctness of our results.

5.1. System with arrivals in every slot: the case $C(0) = 0$

Consider the special case where at least one arrival occurs in every slot on the time axis, i.e., where the probability of having no arrivals in a slot is equal to zero: $C(0) = 0$. In these circumstances, the number of arrivals in the kth time slot can be expressed as $c_k = 1 + e_k$, where we label the component 1 as the "persistent" arrival, and we refer to the random variable e_k as the number of "excess" arrivals in slot k. The pgf $C(z)$ and its derivatives at $z = 1$ can then be expressed as

$$C(z) = zE(z) \quad ; \quad C'(1) = \lambda = 1 + E'(1) = 1 + \lambda_e \quad ; \quad C''(1) = E''(1) + 2\lambda_e \ , \quad (105)$$

where $E(z)$ and λ_e denote the pgf and the expected value of the number of excess arrivals in a slot. One possible view of such a system is as follows: the regular server, which is permanently available, suffices to take care of the persistent arrival stream, removing every persistent arrival from the system one slot after it entered the system, whereas the extra server processes the excess arrivals. From the point of view of the excess arrivals, they have at their disposal one server, the extra server, which is available intermittently, having geometrically distributed up-periods (according to (3)) and arbitrarily distributed down-periods (according to (5)). The total number of customers in the system should therefore, at any slot boundary, be equal to one more than the number of excess customers in the system, the persistent arrivals — except for the one that entered in the last slot before the considered slot boundary — having been also persistently removed from the system by the regular server. We thus expect the important pgfs of the total system content to take the form

$$G_0(z) = zG_{0,e}(z) \quad ; \quad H_0(z) = zH_{0,e}(z) \quad ; \quad U(z) = zU_e(z) \quad , \quad (106)$$

where the functions $G_{0,e}(z)$, $H_{0,e}(z)$ and $U_e(z)$ indicate the pgfs of the system content in a single-server system with only excess arrivals, at the beginning of an up-period, at the beginning of a down-period, and at an arbitrary slot boundary, respectively. As mentioned in subsection 1.2.2, such a single-server system with server interruptions has been analyzed extensively in queueing literature, for instance, in [49, 50, 52, 64]. We stress that equations (106) have been derived without the restriction that the pgf $R(z)$ should be rational, just as in [49, 50, 52, 64], and that the expressions that can be derived by combining (106) with our equations (30), (31) and (43) are basically identical with the findings there.

5.2. System without arrivals: the case $C(z) = 1$

If there are no arrivals in the system, we expect the steady-state number of customers in the system to be deterministically equal to zero, i.e., if $C(z) = 1$, for all z, we expect that also $G_0(z) = 1$, $H_0(z) = 1$ and $U(z) = 1$, for all z. Moreover, this should be true for all possible choices of the server interruption process, i.e., all possible choices of $(R(z), \alpha)$, or, equivalently, $F(z)$. As a consequence, we also expect $E[g_0] = 0$, $E[h_0] = 0$ and $E[u] = 0$, regardless of $F(z)$. Although this seems obvious, our formulas for the pgfs and the expected values of the system content, derived earlier, do not reveal this to be the case without further inspection. Nevertheless we have been able to prove – at the expense of rather lengthy calculations – that our formulas do reduce to the correct results.

6. The limiting cases $\sigma \to 0$ and $\sigma \to 1$

As mentioned in section 2, the parameter σ, defined in (6), can only take values strictly greater than zero and strictly lower than 1. Nevertheless, the cases $\sigma = 0$ and $\sigma = 1$ are interesting in their own right, although - strictly speaking - they are not special cases of our current model. In fact, in both cases, the alternating sequence of up-periods and down-periods no longer exists, and all slots are either down-slots or up-slots. In these circumstances, it does not make sense anymore to study the pgfs $G_k(z)$ and $H_k(z)$ of the system content after the kth slot ($k \geq 1$) of an up-period or a down-period, respectively. The same remark holds for the related pgfs $G_0(z)$, $H_0(z)$, $G(z)$ and $H(z)$, defined in subsection 3.2, but the pgf $U(z)$ of the system content at the beginning of a random slot, first computed in subsection 3.5, remains a meaningful concept. In this section, we therefore focus on the pgf $U(z)$ for these two cases, which are well-known in queueing literature, and also examine to what extent these results can also be obtained by considering the limits $\sigma \to 0$ and $\sigma \to 1$ in our earlier findings.

6.1. The limiting case $\sigma \to 0$

In case $\sigma = 0$, our system reduces to a single-server queue with deterministic 1-slot service times, where the server is permanently available. This kind of discrete-time queueing system has been studied very frequently in the queueing literature (see, e.g. [45, 66, 56, 52, 64, 67]). The main results are

$$U(z) = U_0(z) \triangleq \frac{(1-\lambda)(z-1)C(z)}{z - C(z)} \tag{107}$$

and

$$E[u] = E[u]_0 \triangleq \lambda + \frac{C''(1)}{2(1-\lambda)} \ . \tag{108}$$

It is not difficult to see that exactly the same results are retrieved from the formulas (42) or (43) for $U(z)$, and (44) for $E[u]$, when σ is equated to zero.

6.2. The limiting case $\sigma \to 1$

In case $\sigma = 1$, our system reduces to a two-server queue with deterministic 1-slot service times, where both servers are permanently available. This kind of discrete-time queueing system has also been studied quite extensively in the queueing literature (see, e.g. [66, 56, 64, 67, 19, 23]). In this case, the main results are

$$U(z) = U_1(z) \triangleq \frac{(2-\lambda)(z-1)C(z)}{z^2 - C(z)} \cdot \frac{z - \hat{z}}{1 - \hat{z}} \tag{109}$$

and

$$E[u] = E[u]_1 \triangleq \lambda + \frac{C''(1) - 2}{2(2-\lambda)} + \frac{1}{1-\hat{z}} \ , \tag{110}$$

where the quantity $\hat{z}$ is defined as the only zero of $U_1(z)$'s denominator $z^2 - C(z)$ strictly inside the unit disk of the complex z-plane.

Keeping in mind that $\sigma \to 1$ implies $\alpha \to 1$, we have been able to show that the same results can also be retrieved from our earlier formulas.

7. Examples and numerical results

In this section, we illustrate the developed methodology by means of an extended special case. Specifically, we assume that the down-periods are distributed according to a mixture of two geometrics with respective mean values $r_1 \geq 1$ and $r_2 \geq 1$, and respective weighing coefficients ω and $1-\omega$ $(0 \leq \omega \leq 1)$, i.e., with pmf

$$r(n) = \frac{\omega}{r_1}\big(1-\frac{1}{r_1}\big)^{n-1} + \frac{1-\omega}{r_2}\big(1-\frac{1}{r_2}\big)^{n-1} \quad , \quad n \geq 1 \ , \tag{111}$$

and mean value

$$\bar{r} \triangleq \sum_{n=1}^{\infty} n r(n) = \omega r_1 + (1-\omega) r_2 \geq 1 \ . \tag{112}$$

In order to limit the number of parameters and keep a good amount of flexibility, we choose the quantities ω, r_1 and r_2 in such a way that both geometrics contribute equally to the mean value of the down-periods, i.e.,

$$\omega r_1 = (1-\omega) r_2 = \frac{\bar{r}}{2} \ . \tag{113}$$

Denoting the second moment of the down-periods as $E\big[r^2\big]$, the variance of the down-periods is given by

$$\mathrm{var}[r] = E\big[r^2\big] - (E[r])^2 = \omega r_1^2\big(2-\frac{1}{r_1}\big) + (1-\omega) r_2^2\big(2-\frac{1}{r_2}\big) - \bar{r}^2 \ .$$

In view of (112) and (113), this can be expressed in terms of ω and $\bar{r}$ as follows:

$$\mathrm{var}[r] = 2\omega r_1^2 + 2(1-\omega) r_2^2 - \bar{r} - \bar{r}^2 = \frac{\bar{r}^2}{2}\big(\frac{1}{\omega} + \frac{1}{1-\omega}\big) - \bar{r} - \bar{r}^2 \ ,$$

so that the squared coefficient of variation of the down-periods equals

$$C_r^2 \triangleq \frac{\mathrm{var}[r]}{\bar{r}^2} = \frac{1}{2\omega(1-\omega)} - 1 - \frac{1}{\bar{r}} \ . \tag{114}$$

Equation (114) shows that, with a proper choice of the weighing probability ω between 0 and 1, our mixture of geometric distributions can exhibit any value of the squared coefficient of variation C_r^2 between $1 - 1/\bar{r}$ (for $\omega = 0.5$) and infinity (for $\omega = 0$ or $\omega = 1$).

The pgf $R(z)$ of the down-periods can now be computed from (111), (113) and (114) as

$$R(z) \triangleq \sum_{n=1}^{\infty} r(n) z^n = \frac{2z[z-(1+\bar{r}C_r^2)(z-1)]}{2z^2 - (1+\bar{r}+\bar{r}C_r^2)(z-1)[2z-\bar{r}(z-1)]} \ ,$$

which expresses $R(z)$ completely in terms of the mean value and the squared coefficient of variation of the down-periods, which are parameters with practical significance. It is clear that the pgf $R(z)$ is a rational function of z indeed, so that the methodology

developed in section 4 can be applied. Using the notations introduced in subsection 4.1, we obtain the following results in this specific case:

$$R(z) = \frac{A(z)}{B(z)} \quad , \quad \text{with } A(1) = B(1) = 1 \ ,$$

so that the polynomials $A(z)$ and $B(z)$, defined in (47), are given by

$$A(z) = z[z - (1 + \bar{r}C_r^2)(z-1)] \tag{115}$$

and

$$B(z) = z^2 - \frac{1}{2}(1 + \bar{r} + \bar{r}C_r^2)(z-1)[2z - \bar{r}(z-1)] \ . \tag{116}$$

The mean availability σ of the extra server, defined in (6), is given by

$$\sigma = \frac{1}{1 + (1-\alpha)\bar{r}} \ ,$$

so that the parameter α can be expressed in terms of σ and $\bar{r}$ (which have a greater practical significance) as

$$\alpha = 1 - \frac{1-\sigma}{\sigma\bar{r}} \quad , \quad 1 - \alpha = \frac{1-\sigma}{\sigma\bar{r}} \ . \tag{117}$$

We note that the condition that the mean down-period $1/(1-\alpha)$ is at least equal to 1 slot, implies that

$$\bar{r} \geq \frac{1-\sigma}{\sigma} \ ,$$

or, equivalently,

$$\sigma \geq \frac{1}{1+\bar{r}} \ .$$

The pgf $F(z)$, defined in (7), now takes the form

$$F(z) = z\left(1 + \frac{(1-\sigma)(z-1)[z - \frac{1}{2}(1 + \bar{r} + \bar{r}C_r^2)(z-1)]}{\sigma\{z^2 - \frac{1}{2}(1 + \bar{r} + \bar{r}C_r^2)(z-1)[2z - \bar{r}(z-1)]\}}\right) \ , \tag{118}$$

whereas the quantity $F''(1)$, given in (9), reduces to

$$F''(1) = \frac{1-\sigma}{\sigma}(1 + \bar{r} + \bar{r}C_r^2) \ . \tag{119}$$

Note that the server-interruption process, which – in general – is fully determined by the pgf $F(z)$ of the inter-extra times, has now been completely specified in terms of just three parameters with direct physical meaning, i.e., σ, the fraction of time when the extra server is available, and $\bar{r}$ and C_r^2, the mean value and the squared coefficient of variation of the lengths of the down-periods. These parameters can be varied nearly arbitrarily, with the following restrictions:

$$\bar{r} \geq 1 \quad , \quad C_r^2 \geq 1 - \frac{1}{\bar{r}} \quad , \quad \sigma \geq \frac{1}{1+\bar{r}} \ . \tag{120}$$

The only remaining quantities to be chosen or computed are the pgf $C(z)$ of the arrival process and the zeroes $\{x_1, x_2, x_3\}$ which also appear in the formulas (100), (101), and (102). As a specific example, let us make the (very common) assumption that the number of arrivals per slot has a Poisson distribution with mean λ, i.e.,

$$C(z) = e^{\lambda(z-1)} \ , \tag{121}$$

which implies that the quantity $C''(1)$, appearing in (100), (101), and (102), is given by

$$C''(1) = \lambda^2 \ . \tag{122}$$

According to equation (104), the zeroes $\{x_i, 1 \leq i \leq 3\}$ can be expressed as

$$x_i = \frac{z_i}{C(z_i)} = z_i e^{\lambda(1-z_i)} \quad , \quad 1 \leq i \leq 3 \ , \tag{123}$$

where the quantities $\{z_i, 1 \leq i \leq 3\}$ are the three roots of the equation $z = \hat{F}(z)$, strictly inside the (open) unit disk $\{z : |z| < 1\}$ of the complex z-plane. Here

$$\hat{F}(z) = F\left(\frac{C(z)}{z}\right) = F\big(z^{-1}e^{\lambda(z-1)}\big) \ , \tag{124}$$

where $F(z)$ is given explicitly in equation (118). The equation to solve thus takes the form

$$z^2 = e^{\lambda(z-1)}\left(1+\frac{(1-\sigma)(e^{\lambda(z-1)}-z)[e^{\lambda(z-1)}-\frac{1}{2}(1+\bar{r}+\bar{r}C_r^2)(e^{\lambda(z-1)}-z)]}{\sigma\{e^{2\lambda(z-1)}-\frac{1}{2}(1+\bar{r}+\bar{r}C_r^2)(e^{\lambda(z-1)}-z)[2e^{\lambda(z-1)}-\bar{r}(e^{\lambda(z-1)}-z)]\}}\right) \ . \tag{125}$$

Once the values of $\{z_1, z_2, z_3\}$ have been determined (numerically) from equation (125), the expected values $E[g_0]$, $E[h_0]$ and $E[u]$ can be derived from the formulas (100), (101), and (102).

Some numerical results are depicted graphically in Figs. 2, 3 and 4. Specifically, Fig. 2 shows the mean system content at three different types of observation epochs, namely at the beginning of an up-period ($E[g_0]$), at the beginning of a down-period ($E[h_0]$), and at random slot boundaries ($E[u]$), versus the mean arrival rate λ, for fixed values of the down-period moments ($\bar{r} = 10$ and $C_r^2 = 1$) and two different values of σ ($\sigma = 0.3$ and $\sigma = 0.9$). Various observations can be made. First, the mean availability σ of the extra server has a severe impact on the average number of customers in the system: higher σ implies lower mean system content, as expected. Next, the curves for $\sigma = 0.3$ all have a vertical asymptote at $\lambda = 1 + \sigma = 1.3$, whereas the curves for $\sigma = 0.9$ all have a vertical asymptote at $\lambda = 1 + \sigma = 1.9$, in accordance with the stability condition (13). Finally, the figure also makes clear that, regardless of the value of σ, the three mean system contents can be ordered as

$$E[h_0] < E[u] < E[g_0] \quad , \quad \text{for all values of } \lambda < 1 + \sigma \ , \tag{126}$$

which is consistent with the intuition that the mean system content should be highest just after a down-period (i.e., at the beginning of an up-period: $E[g_0]$), smallest just after an up-period (i.e., at the beginning of a down-period: $E[h_0]$), and somewhere in between at random slot boundaries ($E[u]$).

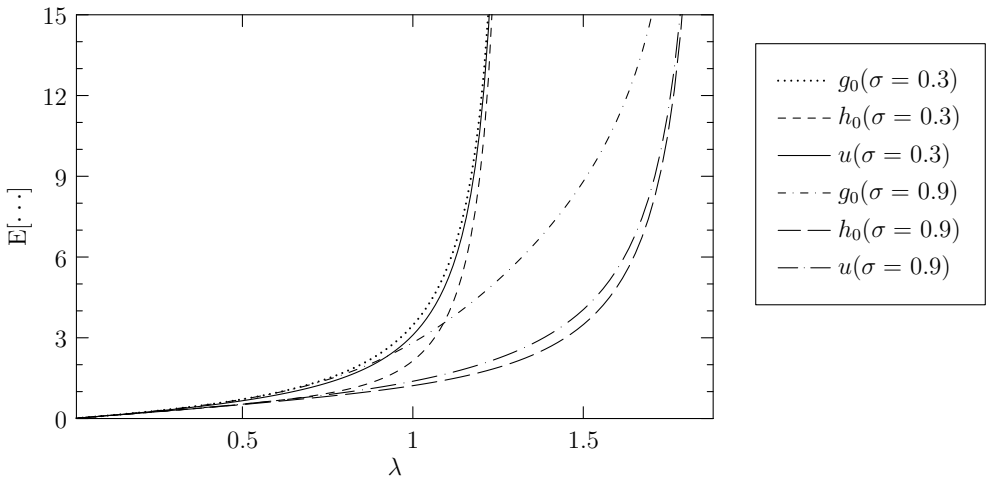

Figure 2: Mean system contents at the beginning of an up-period ($E[g_0]$), at the beginning of a down-period ($E[h_0]$), and at random slot boundaries ($E[u]$) versus mean arrival rate λ, for $\bar{r} = 10$, $C_r^2 = 1$ and two values of σ.

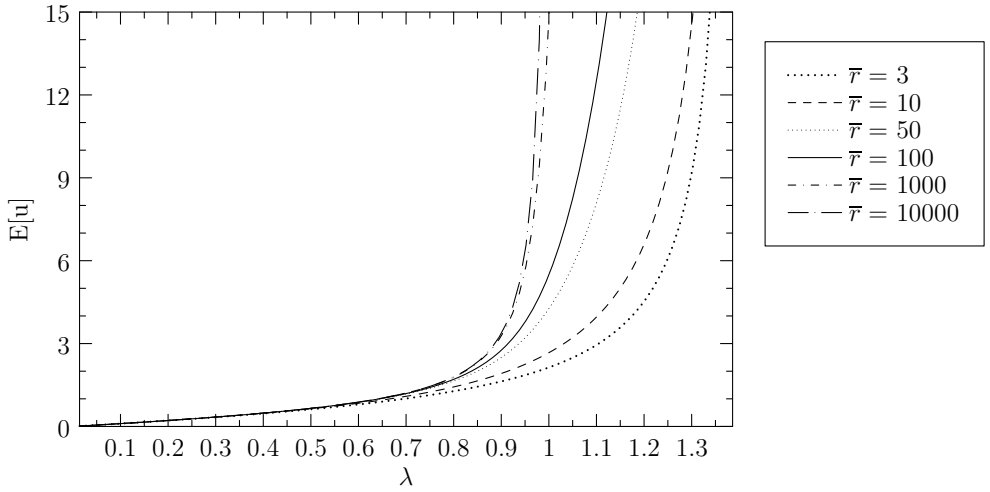

Figure 3: Mean system content at random slot boundaries ($E[u]$) versus mean arrival rate λ, for $\sigma = 0.4$, $C_r^2 = 1$ and various values of $\bar{r}$.

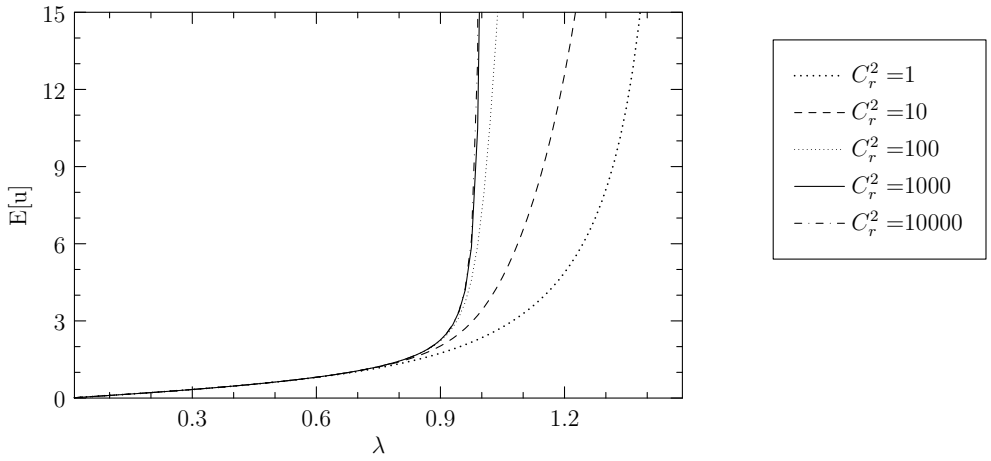

Figure 4: Mean system content at random slot boundaries ($E[u]$) versus mean arrival rate λ, for $\sigma = 0.5$, $\bar{r} = 10$ and various values of C_r^2.

Fig. 3 focuses on the influence of the absolute (mean) lengths of the up-periods and the down-periods for a given ratio of these mean periods, or, equivalently, for a given mean availability σ of the extra server. In particular, Fig. 3 shows the mean system content at random slot boundaries ($E[u]$) versus the mean arrival rate λ, for a fixed value of $\sigma = 0.4$; the absolute mean lengths of the up-periods and the down-periods, given by $2\bar{r}/3$ and $\bar{r}$ respectively in this case, are varied by choosing various values for the parameter $\bar{r}$ as indicated; the squared coefficient of variation of the down-periods is kept constant at $C_r^2 = 1$. The figure clearly shows that, for given (λ and) σ, the mean system content increases considerably with $\bar{r}$, unless the mean arrival rate is relatively small and, hence, the presence or absence of the extra server is not very important. This means that, for large enough arrival loads, not only the fraction of time (σ) the extra server is available matters but also the absolute mean durations of the availability and unavailability periods. A similar phenomenon was observed in the related paper [59], as well as in various other studies of queues with server interruptions.

In Fig. 4, we examine the influence of the variability of the down-periods (as characterized by their squared coefficient of variation C_r^2), once the absolute mean lengths of up-periods and down-periods have been fixed. Specifically, Fig. 4 shows the mean system content at random slot boundaries ($E[u]$) versus the mean arrival rate λ, for fixed values of $\sigma = 0.5$ and $\bar{r} = 10$, for various values of C_r^2 as indicated. All the curves in the figure have a vertical asymptote at $\lambda = 1 + \sigma = 1.5$, as expected. Again, we observe that for low arrival rates, the effect of the variability of the down-periods on the queueing performance is not very important, but for higher arrival rates, the impact of that variability is detrimental, even for given mean up-periods and down-periods of the extra server. Intuitively, higher values of C_r^2 entail the occasional occurrence of very long down-periods, resulting in temporary high accumulations of customers in the system, which also remain in the system for long periods of time.

8. Conclusions and future work

This paper has considered a discrete-time queueing model with general independent arrivals and two identical servers with deterministic service times equal to 1 time slot. One server is permanently available while the other is subject to external random interruptions, characterized by an alternating sequence of random-length up-periods and down-periods. We have been able to compute the pgfs and the expected values of the number of customers in the system at various observation epochs of practical interest, under the restriction that the up-periods be geometrically distributed and that the down-periods have a rational pgf. All the results are semi-analytical, in the sense that they are expressed in terms of the original model parameters, on the one hand, and a finite number of roots of a known non-linear equation, on the other hand. Various models analyzed in earlier papers can be obtained as special cases of the current model, and we have been able to observe that the results are consistent. An interesting special case that was not reported before in the literature (down-periods distributed according to a mixture of two geometric distributions) was analyzed in great detail. With this particular model, the mean and the coefficient of variation of the down-periods can be varied largely independently, which has enabled us to examine the impact of these quantities on the queueing performance.

A major part of the paper was devoted to the detailed study of the unknown parameters appearing in the initial expressions (30), (31) and (43) of the system-content pgfs $G_0(z)$, $H_0(z)$ and $U(z)$, i.e., the expressions obtained after mechanically solving the steady-state equations, before invoking the boundedness of pgfs inside the closed unit disk of the complex plane. These unknown parameters, i.e., the quantities $p(0)$ and $p(1)$, defined in (27), and $q(i), i \geq 1$, defined in (24), are, in fact, closely related to the probability that the service capacity is not being fully used, either during up-periods or down-periods. Specifically, $p(0)$ and $p(1)$ have to do with the probability of having less than two customers in the system, during up-periods (when two servers are available), whereas $q(i), i \geq 1$ similarly relate to the probability of having less than one customer in the system, i.e., the probability of an empty system, during down-periods (when one server is available). The fact that only two unknown parameters ($p(0)$ and $p(1)$) turn up for the up-periods is a consequence of the geometric nature of the up-period distribution. From our results it is clear that, for the down-periods, just one single additional unknown parameter (say $\hat{p}(0)$) shows up in case the down-period distribution is also geometric, meaning that the unknowns $q(i), i \geq 1$ can all be expressed in terms of $\hat{p}(0)$ in that case. Similarly, it can be observed that the down-periods bring about $\hat{r}$ additional independent unknowns if the down-period distribution is deterministic with value $\hat{r}$. Finally, in the (so far) most general case where the down-period pgf $R(z)$ is a ratio of two polynomial functions, the number of additional independent unknowns associated with the down-periods is equal to $\hat{m} \triangleq \max\{m_A, m_B\}$, where m_A and m_B denote the degrees of the numerator and denominator polynomial of $R(z)$, respectively. Although we have no formal proof of this, we strongly conjecture that the number of independent unknown parameters related to the down-periods is infinitely large in all other cases, i.e., for all non-rational pgfs $R(z)$.

A further observation we make is that the (unfortunate) circumstance that both the up-periods and the down-periods of the server-interruption process give rise to a number of unknown parameters in the initial expressions for the system-content pgfs, is due to

the fact that the number of available servers in the system under study is always strictly positive. This gives rise to the presence of non-linear terms of the form $(u_{k-1} - m)^+$, where $m > 0$ denotes the number of available servers, in the system equations for the number of customers (say u_k at the start of slot k) in the system, i.e., in this case, the equations (14) and (15). This also explains why the analysis of the current model is much more complicated than for many earlier models where the number of available servers could also be zero during certain time periods: during such periods, $m = 0$ and the non-linearity in the system equation, associated with the $(\cdots)^+$-operator, disappears, i.e., no additional unknown parameters emerge.

Future work could incorporate more general distributions for the up-periods than the geometric distribution considered here. We may expect that, for the up-periods, extensions will be possible from the geometric distribution to more general distributions with rational pgf, along the same lines as described for the down-periods in the current paper. We strongly suspect that the number of unknown parameters associated with the up-periods will, however, basically be "twice as high", because the number of available servers is two instead of one. Other possible extensions could consider systems with more than two servers, more general service-time distributions, finite waiting rooms, time-correlated arrival processes, etc. Also, the derivation of the full distribution (or pgf) of customer delays and waiting times could be envisaged.

Acknowledgement

This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (BelSPO), Belgium.

References

[1] P. Hokstad, Steady-state solution of the M/G/2 queue, Advances in Applied Probability 11 (1) (1979) 240–255.

[2] J. Cohen, On the M/G/2 queueing model, Stochastic Processes and their Applications 12 (1982) 231-248.

[3] C. Knessl, B. Matkowsky, Z. Schuss, C. Tier, An integral-equation approach to the M/G/2 queue, Operations Research 38 (3) (1990) 506–518.

[4] G. Koole, A simple proof of the optimality of a threshold policy in a two-server queueing system, Systems & Control Letters 26 (5) (1995) 301–303.

[5] O. Boxma, Q. Deng, A. Zwart, Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers, Queueing Systems 40 (1) (2002) 5–31.

[6] M. Van Hoorn, H. Tijms, Approximations for the waiting time distribution of the M/G/C queue, Performance Evaluation 2 (1) (1982) 22–28.

[7] J. De Smit, The queue GI/M/s with customers of different types or the queue GI/H_m/s, Advances in Applied Probability 15 (2) (1983) 392-419.

[8] D. Bertsimas, An analytical approach to a general class of G/G/s queueing systems, Operations Research 38-(1) (1990) 139-155.

[9] R. Chakka, T. Do, The MM $\sum_{k=1}^{K}$ CPP$_k$/GE/c/L G-queue with heterogeneous servers: Steady state solution and an application to performance evaluation, Performance Evaluation 64 (3) (2007) 191–209.

[10] A. Brandwajn, T. Begin, Reduced complexity in M/Ph/c/N queues, Performance Evaluation 78 (2014) 42–54.

[11] C. Kim, A. Dudin, S. Dudin, O. Dudina, Hysteresis control by the number of active servers in queueing system MMAP/PH/N with priority service, Performance Evaluation 101 (2016) 20–33.

[12] D. Lee, S. Li, Transient analysis of multiserver queues with Markov-Modulated Poisson arrivals and overload control, Performance Evaluation 16 (1-3) (1992) 49–66.

[13] S. Chakravarthy, H. Karatza, Two-server parallel system with pure space sharing and Markovian arrivals, Computers & Operations Research 40 (1) (2013) 510–519.
[14] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys, J. Walraevens, A continuous-time queueing model with class clustering and global FCFS service discipline, Journal of Industrial and Management Optimization 10 (2014) 193–206.
[15] B. Kim, J. Kim, Stability of a two-class two-server retrial queueing system, Performance Evaluation 88-89 (2015) 1–17.
[16] W. Mélange, J. Walraevens, D. Claeys, B. Steyaert, H. Bruneel, The impact of a global FCFS service discipline in a two-class queue with dedicated servers, Computers & Operations Research 71 (2016) 23–33.
[17] H. Bruneel, B. Steyaert, Buffer requirements for ATM switches with multiserver output queues, Electronics Letters 27 (8) (1991) 671–673.
[18] H. Bruneel, B. Steyaert, E. Desmet, G. Petit, 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.
[19] H. Bruneel, B. Steyaert, E. Desmet, G. Petit, Analytic derivation of tail probabilities for queue lengths and waiting-times in ATM multiserver queues, European Journal of Operational Research 76 (3) (1994) 563–572.
[20] K. Sohraby, J. Zhang, Spectral decomposition approach for transient analysis of multiserver discrete-time queues, Performance Evaluation 21 (1-2) (1994) 131–150.
[21] N. Kim, M. Chaudhry, B. Yoon, K. Kim, A complete and simple solution to a discrete-time finite-capacity BMAP/D/c queue, Applied Mathematics 3 (12A) (2012) 2169–2173.
[22] B. Vinck, H. Bruneel, Delay analysis of multiserver ATM buffers, Electronics Letters 32 (15) (1996) 1352–1353.
[23] H. Bruneel, I. Wuyts, Analysis of discrete-time multiserver queuing models with constant service times, Operations Research Letters 15 (5) (1994) 231–236.
[24] P. Gao, S. Wittevrongel, H. Bruneel, Delay analysis for a discrete-time GI-D-c queue with arbitrary-length service times, Lecture Notes in Computer Science 3236 (2004) 184–195.
[25] N. Kim, M. Chaudhry, The use of the distributional Little's law in the computational analysis of discrete-time GI/G/1 and GI/D/c queues, Performance Evaluation 65 (1) (2008) 3–9.
[26] P. Gao, S. Wittevrongel, J. Walraevens, H. Bruneel, Analytic study of multiserver buffers with two-state Markovian arrivals and constant service times of multiple slots, Mathematical Methods of Operations Research 67 (2) (2008) 269–284.
[27] P. Gao, S. Wittevrongel, H. Bruneel, Discrete-time multiserver queues with geometric service times, Computers & Operations Research 31 (1) (2004) 81–99.
[28] P. Gao, S. Wittevrongel, H. Bruneel, Delay against system contents in discrete-time G/Geom/c queue, Electronics Letters 39 (17) (2003) 1290–1292.
[29] P. Gao, S. Wittevrongel, H. Bruneel, On the behavior of multiserver buffers with geometric service times and bursty input traffic, IEICE Transactions on Communications E87B (12) (2004) 3576–3583.
[30] V. Goswami, Analysis of discrete-time multi-server queue with balking, International Journal of Management Science and Engineering Management 9 (1) (2014) 21–32.
[31] Q. He, A. Alfa, Construction of Markov chains for discrete time MAP/PH/K queues, Performance Evaluation 93 (2015) 17–26.
[32] A. Krishnamoorthy, P. Pramod, S. Chakravarthy, Queues with interruptions: a survey, TOP 22 (1) (2014) 290–320.
[33] D. Gaver, A waiting line with interrupted service, including priorities, Journal of the Royal Statistical Society. Series B (Methodological) 24 (1) (1962) 73–90.
[34] I. Sahin, Single server queue with preemptive service interruptions, Journal of Applied Probability 8 (4) (1971) 835–837.
[35] I. Mitrany, B. Avi-Itzhak, A many-server queue with service interruptions, Operations Research 16 (3) (1968) 628–638.
[36] P. Nain, Queuing-systems with service interruptions - an approximation model, Performance Evaluation 3 (2) (1983) 123–129.
[37] N. Van Dijk, Simple bounds for queuing-systems with breakdowns, Performance Evaluation 8 (2) (1988) 117–128.
[38] F. Hillier, K. So, The assignment of extra servers to stations in tandem queuing-systems with small or no buffers, Performance Evaluation 10 (3) (1989) 219–231.
[39] O. Ibe, J. Keilson, Multiserver threshold queues with hysteresis, Performance Evaluation 21 (3)

(1995) 185–213.
[40] J. Lui, L. Golubchik, Stochastic complement analysis of multi-server threshold queues with hysteresis, Performance Evaluation 35 (1-2) (1999) 19–48.
[41] M. Jain, P. Singh, Performance prediction of loss and delay Markovian queueing model with nopassing and removable additional servers, Computers & Operations Research 30 (8) (2003) 1233–1253.
[42] Z. Zhang, Performance analysis of a queue with congestion-based staffing policy, Management Science 55 (2) (2009) 240–251.
[43] E. Ruelas-Gonzalez, J. Limon-Robles, N. Smith-Cornejo, Determining a checkout register opening policy to maximize profit in convenience stores chains, Journal of Applied Research and Technology 8 (3) (2010) 406–415.
[44] C. Kim, A. Dudin, Analysis of a queueing model with contingent additional server, in: International Conference on Computer Networks, Springer, 2016, pp. 306–315.
[45] J. Hsu, Buffer behavior with Poisson arrival and geometric output processes, IEEE Transactions on Communications 22 (1974) 1940–1941.
[46] T. Heines, Buffer behavior in computer communication systems, IEEE Transactions on Communications 28 (1979) 573–576.
[47] N. Georganas, Buffer behavior with Poisson arrivals and bulk geometric output processes, IEEE Transactions on Communications 24 (8) (1976) 938–940.
[48] D. Towsley, The analysis of a statistical multiplexer with nonindependent arrivals and errors, IEEE Transactions on Communications 28 (1) (1980) 65–72.
[49] H. Bruneel, Analysis of buffer behavior for an integrated voice-data system, Electronics Letters 19 (2) (1983) 72–74.
[50] H. Bruneel, On the behavior of buffers with random server interruptions, Performance Evaluation 3 (3) (1983) 165–175.
[51] H. Bruneel, Analysis of an infinite buffer system with random server interruptions, Computers & Operations Research 11 (4) (1984) 373–386.
[52] H. Bruneel, A general treatment of discrete-time buffers with one randomly interrupted output line, European Journal of Operational Research 27 (1) (1986) 67–81.
[53] C. Woodside, E. Ho, Engineering calculation of overflow probabilities in buffers with Markov-interrupted service, IEEE Transactions on Communications 35 (12) (1987) 1272–1277.
[54] D. Lee, Analysis of a single server queue with semi-Markovian service interruption, Queueing Systems 27 (1-2) (1997) 153–178.
[55] M. Ali, X. Zhang, J. Hayes, A performance analysis of a discrete-time queueing system with server interruption for modeling wireless ATM multiplexer, Performance Evaluation 51 (1) (2003) 1–31.
[56] H. Bruneel, A general model for the behavior of infinite buffers with periodic service opportunities, European Journal of Operational Research 16 (1) (1984) 98–106.
[57] K. Laevens, H. Bruneel, Discrete-time multiserver queues with priorities, Performance Evaluation 33 (4) (1998) 249–275.
[58] B. Vinck, H. Bruneel, System delay versus system content for discrete-time queueing systems subject to server interruptions, European Journal of Operational Research 175 (1) (2006) 362–375.
[59] H. Bruneel, A mathematical model for discrete-time buffer systems with correlated output process, European Journal of Operational Research 18 (1) (1984) 98–110.
[60] H. Bruneel, A discrete-time queueing system with a stochastic number of servers subjected to random interruptions, Opsearch 22 (4) (1985) 215–231.
[61] D. Fiems, B. Steyaert, H. Bruneel, Discrete-time queues with generally distributed service times and renewal-type server interruptions, Performance Evaluation 55 (3-4) (2004) 277–298.
[62] T. Takine, B. Sengupta, A single server queue with service interruptions, Queueing Systems 26 (3-4) (1997) 285–300.
[63] E. Morozov, D. Fiems, H. Bruneel, Stability analysis of multiserver discrete-time queueing systems with renewal-type server interruptions, Performance Evaluation 68 (12) (2011) 1261–1275.
[64] H. Bruneel, B. Kim, Discrete-time models for communication systems including ATM, Kluwer Academic, Boston, USA, 1993.
[65] M. Gonzáles, Classical complex analysis, Marcel Dekker, New York, USA, 1992.
[66] J. Hunter, Discrete-time models: techniques and applications, Academic Press, New York, 1983.
[67] H. Takagi, Queueing Analysis; A foundation of performance evaluation, volume 3: Discrete-time systems, Elsevier Science Publishers, Amsterdam, 1993.