

Analysis of a two-class single-server discrete-time FCFS queue: the effect of interclass correlation

Herwig Bruneel*, Tom Maertens*, Bart Steyaert*, Dieter Claeys*⁺,
Dieter Fiems*, Joris Walraevens*

*SMACS Research Group

Dept. of Telecommunications and Information Processing

⁺Dept. of Industrial Systems Engineering and Product Design

Ghent University - UGent

Abstract

In this paper we study a discrete-time queueing system with one server and two classes of customers. Customers enter the system according to a general independent arrival process. The classes of consecutive customers, however, are correlated in a Markovian way. The system uses a “global FCFS” service discipline, i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their classes. The service-time distribution of the customers is general but class-dependent, and therefore, the exact order in which the customers of both classes succeed each other in the arrival stream is important, which is reflected by the complexity of the system content and waiting time analysis presented in this paper. In particular, a detailed waiting time analysis of this kind of multiclass system has not yet been published, and is considered to be one of the main novelties by the authors.

In addition to that, a major aim of the paper is to estimate the impact of *interclass correlation* in the arrival stream on the total number of customers in the system, and the customer delay. The results reveal that the system can exhibit two different classes of stochastic equilibrium: a “strong” equilibrium where both customer classes give rise to stable behavior individually, and a “compensated” equilibrium where one customer type creates overload.

Key words: discrete-time queueing; multi-class; waiting time; interclass correlation; global FCFS; general service times

1 Introduction

Multi-class queueing systems arise when multiple classes (or classes) of customers compete for the use of the same resources. A vast literature on multi-class queueing systems exists, both in a continuous-time setting and in a discrete-time setting [41, 2, 42, 18]. Various classes of scheduling disciplines (i.e., rules that determine the order of service for the customers of different classes) have been investigated. We mention, among others, priority scheduling [16, 35, 25, 48, 19, 43, 45, 32, 44, 49, 2, 42], weighted fair queueing (WFQ) [14, 47, 40, 21], random order of service (ROS) [28, 5], generalized processor sharing (GPS) [46, 37, 5, 27, 36, 26, 30]. Strangely enough, only few results have been derived for multi-class first-come-first-served (*global FCFS*) systems, i.e., queueing systems in which the customers of different classes are accommodated in one single queue and served in their order of arrival, irrespective of the classes they belong to (two recent papers are [16, 17]). The present paper presents the analysis of a discrete-time model that fits in this category.

In multi-class queueing models, it is often assumed that the various customer classes occur randomly and independently in the overall arrival process. (see [16, 17]). In the current paper, however, we explicitly aim at examining the effect of *interclass correlation* (or *class clustering*) in the arrival process, when the queueing discipline is *global FCFS*. In other words, we want to investigate how the performance of a (global FCFS) queueing system depends on the order in which the customers of different classes arrive and are being served. In particular, we are interested to know whether or not the degree to which customers of the same class have the tendency to be served closely together (i.e., back-to-back) has a substantial impact on the performance of a global-FCFS queueing system. Therefore, we superimpose a two-state Markovian model with interclass correlation (and arbitrary transition probabilities for the *classes* of the consecutive customers in the arrival stream) on top of a regular general independent arrival process model for the *numbers* of arrivals from slot to slot. Service-time distributions are general and class-dependent. For this model, we present an extensive analysis of the system content and customer waiting time. Whereas the system content analysis is partly based on the conference paper [9], the customer waiting time analysis is new.

Interclass correlation between consecutive customer classes can be viewed as a form of dependence between the lengths of consecutive customer service times. In this sense there is some resemblance between our model and the so-called Markovian Service Process (MSP) that has been dealt with in a number of earlier papers [1, 38, 4, 23]. Our model is different from the service models used in these papers, in several respects. First of all, our model is a discrete-time model, whereas [1, 38, 4, 23] consider time as a continuous quantity. Next, although both in MSPs and our model the service process is controlled by a Markovian background state, there is a fundamental difference. In our model, the background state is associated with a customer, i.e., it basically represents its class, and can only change from customer to customer. In other words, a next state in our model corresponds to another customer and thus goes along with a service completion. In an MSP on the other hand, the background state is associated with time and can even change during the service of a customer. Ozawa [38] presents a generic matrix-based analysis of a queue where the service process is governed by a background Markov chain having two classes of states (for empty queue and non-empty queue, respectively). In our view, the major contribution of this paper is the derivation of a matrix-class stochastic factorization for the vector generating function of the queue content. Banik et al's paper [4] considers a service process where the consecutive customers are served in batches of random sizes, the consecutive batch sizes (rather than the service times) being correlated. Finally, Gupta et al analyze a model where the service times (or job sizes, as they call them) have a hyperexponential (instead of general) distribution, and the main emphasis is on the behavior

of various non-FCFS scheduling policies.

The current paper is related to a series of papers we published the last couple of years on dual-class global FCFS-queues with interclass correlation in the arrival process [12, 13, 15, 33, 34, 39, 9]. In our conference paper [12] and its extended version [13], we study a system with two *dedicated* servers, i.e., where each server can only serve one class of customers, whereas there is only one server in the current model. In addition, it is assumed in [12, 13] that both customer classes are equiprobable, i.e., both customer classes account for half of the load, whereas this restriction is relaxed in this paper. Also, [12] and [13] consider single-slot service times for both customer classes whereas general service times that may be different for each class are included here. Moreover, the current paper puts a substantial focus on the concepts of global, strong and compensated stability, which are not discussed in [12] and [13]. [15] describes the application of the results from [12] and [13] to the evaluation of in-order processing systems. In [33, 34], a system similar to [12] and [13] is studied though in continuous time instead of discrete time. In [39], a system similar to the one in the current paper is studied, i.e., there are two different classes of customers, the service is granted according to the global FCFS rule, but the service times of the consecutive customers are determined by the equality or non-equality of the classes these customers belong to, rather than by their actual classes. In fact, the current paper is a thoroughly extended version of our conference paper [9]. The main extensions are a more complete discussion of earlier relevant work, a larger number of numerical examples and, especially, the waiting time analysis, and a careful investigation of the system behavior in case of strong positive interclass correlation in section 7, which we believe is a key contribution of this paper as well.

For the current model, we first derive the steady-state probability generating function (pgf) of the total number of customers in the system, at customer departure times. From this result, we then obtain the corresponding pgf valid at arbitrary slot boundaries. Various performance measures of practical use, such as the mean system content (and, from Little's result, the mean customer waiting time and delay), can be easily derived from these pgf's. In addition, we present an extensive analysis that leads to an expression for the steady-state pgf of the customer waiting time. We would like to emphasize that contributions which consider the customer waiting time analysis (other than the mean value) in multiclass systems where customers with class-dependent service times are accommodated by a common queue, are scarce, and to the best of our knowledge, only [1, 17] consider such results. Our mathematical derivations follow a completely different, and in our view novel, approach, and we therefore consider this particular waiting time analysis as another main contribution of the paper.

In addition to that, we show that a close investigation of the resulting mathematical formulas and a number of numerical examples reveals that the system under study can exhibit two classes of stochastic equilibrium, depending on the values of the system parameters: a "strong" regime in which both customer classes individually generate less work than the system can handle (during periods where only such customers arrive), and a "compensated" class of regime whereby one customer class creates overload situations which are compensated by strong underload periods generated by the other customer class. In the latter case, our results clearly demonstrate the crucial importance of the amount of interclass correlation on the usual performance parameters of the system.

Applications of the model are numerous: the two customer classes could model, for instance, two classes of road traffic at a traffic light (see [13] for additional explanation), two classes of

digital information (data, voice, video) in routers of a communication network, two classes of files to be handled by a public servant, etc. One particular application that the authors have encountered in a truck manufacturing plant, is the modeling of the the preassembly process of axles next to the final assembly line of trucks, which is one of the main bottlenecks in the overall production process. Axles of several types are required, corresponding to several types of trucks scheduled in the main assembly production line. Following the just-in-time manufacturing principle, axles therefore have to be preassembled in the order they are needed at the main assembly line, which relates to the global FCFS service order in the model. In addition, the product characteristics, such as whether or not an axle needs air suspension, greatly affects the production time. Therefore, the production times of axles can differ significantly, which is accounted for by the class-dependent service times in our model. In addition, clustering of same-type axles occurs as well, due to the scheduling of the truck production on the main assembly line, and production specifics such as the number of axles needed by the different types of trucks. This clustering effect is captured by the correlation structure of successive types in the model. Due to these features, we believe that the model presented hereafter provides a workable framework for the performance assessment of such a production process.

2 Mathematical model

We consider a discrete-time queueing system with infinite waiting room, one server, and two classes (classes) of customers, named A and B . The time axis is divided into fixed-length intervals referred to as *time slots*, *time units* or, simply, *slots*, in the sequel. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries. Customers are served in their order of arrival, regardless of the class they belong to. We call this service discipline “global FCFS” in this paper.

The arrival process of new customers in the system is characterized in two steps.

First, we model the total (aggregated) arrival stream of new customers by means of a sequence of i.i.d. nonnegative discrete random variables with common probability mass function (pmf) $e(n)$ and common probability generating function (pgf) $E(z)$. More specifically,

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

$$E(z) \triangleq \sum_{n=0}^{\infty} e(n) z^n \quad .$$

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

$$\lambda = E'(1) \quad .$$

Next, we describe the occurrence of the two classes (A and B) in the sequence of the consecutively arriving customers. In this study, we assume that both classes of customers account for part of the total load of the system, i.e., both customer classes are “mixed” in the arrival stream, but there may be some degree of “class clustering” in the arrival process, i.e., customers of any given class may (or may not) have a tendency to “arrive back-to-back”. Mathematically,

fig1-eps-converted-to.pdf

Figure 1: Two-state Markov chain for the A resp. B customer class

this means that the classes of two consecutive customers may be non-independent. Specifically, we assume a first-order Markovian class of correlation between the classes of two consecutive customers, which basically means that the probability that the next customer belongs to a given class depends on the class of the previous customer. Note that our arrival model is profoundly different from classical (discrete-time) correlated arrival models, where the global *numbers* of arrivals are correlated from slot to slot. This is not the case here, but, on the other hand, the *classes* of consecutive customers in the arrival stream are correlated.

Let t_k denote the class (i.e., A or B) of customer k . The transition probabilities of the Markov chain that determines the classes of the consecutive customers are then defined as (see Fig. 1)

$$\begin{aligned} \Pr[t_{k+1} = A | t_k = A] &= \alpha \quad ; \quad \Pr[t_{k+1} = B | t_k = A] = 1 - \alpha \quad , \\ \Pr[t_{k+1} = A | t_k = B] &= 1 - \beta \quad ; \quad \Pr[t_{k+1} = B | t_k = B] = \beta \quad . \end{aligned} \quad (1)$$

It is well-known [8, 6] that for a two-state Markov chain of this class, the steady-state probabilities t_A and t_B of finding the chain in state A or B respectively, are given by

$$\begin{aligned} t_A &\triangleq \lim_{k \rightarrow \infty} \Pr[t_k = A] = \frac{1 - \beta}{2 - \alpha - \beta} \quad , \\ t_B &\triangleq \lim_{k \rightarrow \infty} \Pr[t_k = B] = \frac{1 - \alpha}{2 - \alpha - \beta} \quad . \end{aligned} \quad (2)$$

The quantities t_A and t_B can be interpreted as the fractions of class A and class B customers in the arrival stream, respectively. The (steady-state) correlation coefficient of the Markov chain, i.e., the amount of correlation between the classes of two consecutive customers in the arrival stream (in the steady state), is given by

$$\gamma \triangleq \lim_{k \rightarrow \infty} \frac{E[T_k T_{k+1}] - E[T_k] E[T_{k+1}]}{\sqrt{\text{var}[T_k] \text{var}[T_{k+1}]}} = \alpha + \beta - 1 \quad , \quad (3)$$

where T_k is a numerical random variable corresponding with the non-numerical variable t_k , e.g. $T_k = 1 \Leftrightarrow t_k = A$ and $T_k = 0 \Leftrightarrow t_k = B$. We will indicate the parameter γ ($-1 \leq \gamma \leq +1$) as the *interclass correlation* in the sequel. Positive values of γ correspond to a situation whereby the customers of any given class have a tendency to cluster, while negative values of γ refer to

arrival streams in which the customers of classes A and B have a tendency to alternate, i.e., be mixed more strongly. The case where $\gamma = 0$, of course, corresponds to the classical assumption that classes of subsequent customers are independent.

The service process of the system is characterized by attaching to each customer a corresponding *service requirement* or *service time*, which indicates the number of time slots required to give complete service to the customer at hand. The service times of consecutive customers arriving at the system are modeled as a sequence of independent positive discrete random variables with a class-dependent distribution. More specifically, the pmf's are given by

$$a(n) \triangleq \Pr[\text{service time of class-}A \text{ customer equals } n \text{ slots}] \quad , \quad n \geq 1 \quad ,$$

$$b(n) \triangleq \Pr[\text{service time of class-}B \text{ customer equals } n \text{ slots}] \quad , \quad n \geq 1 \quad ,$$

while the corresponding pgf's are

$$A(z) \triangleq \sum_{n=1}^{\infty} a(n) z^n \quad , \quad B(z) \triangleq \sum_{n=1}^{\infty} b(n) z^n \quad . \quad (4)$$

The mean service times of class- A and class- B customers are given by

$$\mu_A \triangleq A'(1) \quad , \quad \mu_B \triangleq B'(1) \quad . \quad (5)$$

The structure of the rest of this paper is as follows. Section 3 first presents an analysis of the total number of customers in the system at customer departure times: an expression is derived for the pgf of this number and a method is described to determine the two remaining unknowns in that expression; next, the steady-state pgf of the system content at random slot boundaries is derived from this result. All these derivations are valid for arbitrary choices of $E(z)$, $A(z)$, $B(z)$, α and β . Section 4 focuses on the calculation of the steady-state pgf of the customer waiting time. In order to do so, we first need to derive some results on the joint pgf of the system content and the residual service time of the customer being served. Three special cases, whereby the system basically reduces to a single-class model, are considered in Section 5. We discuss the general case, both conceptually and quantitatively, in Section 6. Section 7 focuses on the case of strong positive interclass correlation, i.e., the limit where $\gamma \rightarrow +1$. Some conclusions are drawn and directions for future work are given in Section 8.

3 System content analysis

The results in this section are based on the conference paper [9].

3.1 System equations at customer departure times

Let u_k denote the total *system content*, i.e., the total number of customers present in the system just after the service completion of the k -th customer, and, as before, let t_k indicate the class customer k belongs to. Then, as a consequence of all the model assumptions in section 2, the couple (t_k, u_k) forms a Markovian state description of the system (at customer departure times).

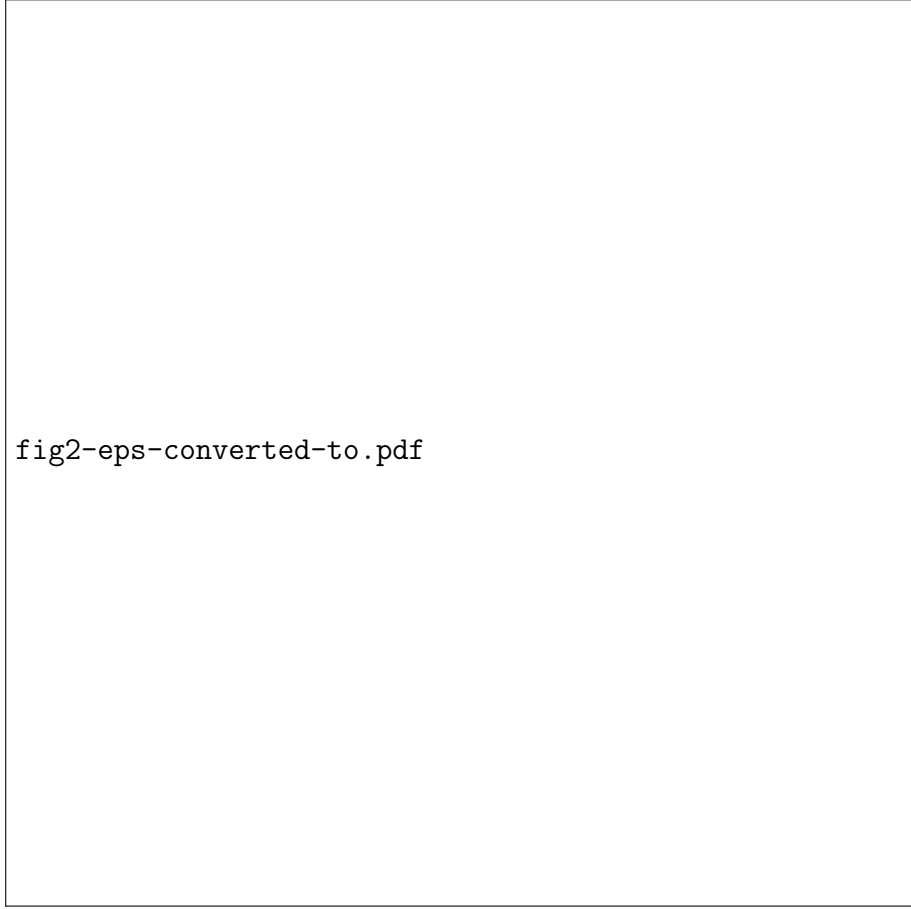


fig2-eps-converted-to.pdf

Figure 2: Relationship between u_k and u_{k+1} when $u_k > 0$

The state transitions of the quantities $\{t_k\}$ are governed by the equations (1), whereas for the quantities $\{u_k\}$, the following recursive system equations can be established (see Figs. 2 and 3):

$$\begin{aligned}
 u_{k+1} &= u_k - 1 + g_{k+1} \quad , \quad \text{if } u_k > 0 \quad , \\
 u_{k+1} &= h_{k+1} \quad , \quad \text{if } u_k = 0 \quad .
 \end{aligned}
 \tag{6}$$

Here, the quantity g_{k+1} is defined as the number of arrivals in the system during the service time of customer $k + 1$, while the quantity h_{k+1} is given by

$$h_{k+1} = g_{k+1} + f_{k+1} \quad ,$$

where f_{k+1} indicates the number of arrivals in the arrival slot of customer $k + 1$ arriving *after* customer $k + 1$ (when customer $k + 1$ arrives in an empty system).

It is easily seen that the pgf of f_{k+1} is given by the pgf of the number of additional arrivals in a slot with at least one arrival, i.e.,

$$F(z) \triangleq E[z^{f_{k+1}}] = \frac{E(z) - E(0)}{z[1 - E(0)]} \quad , \tag{7}$$

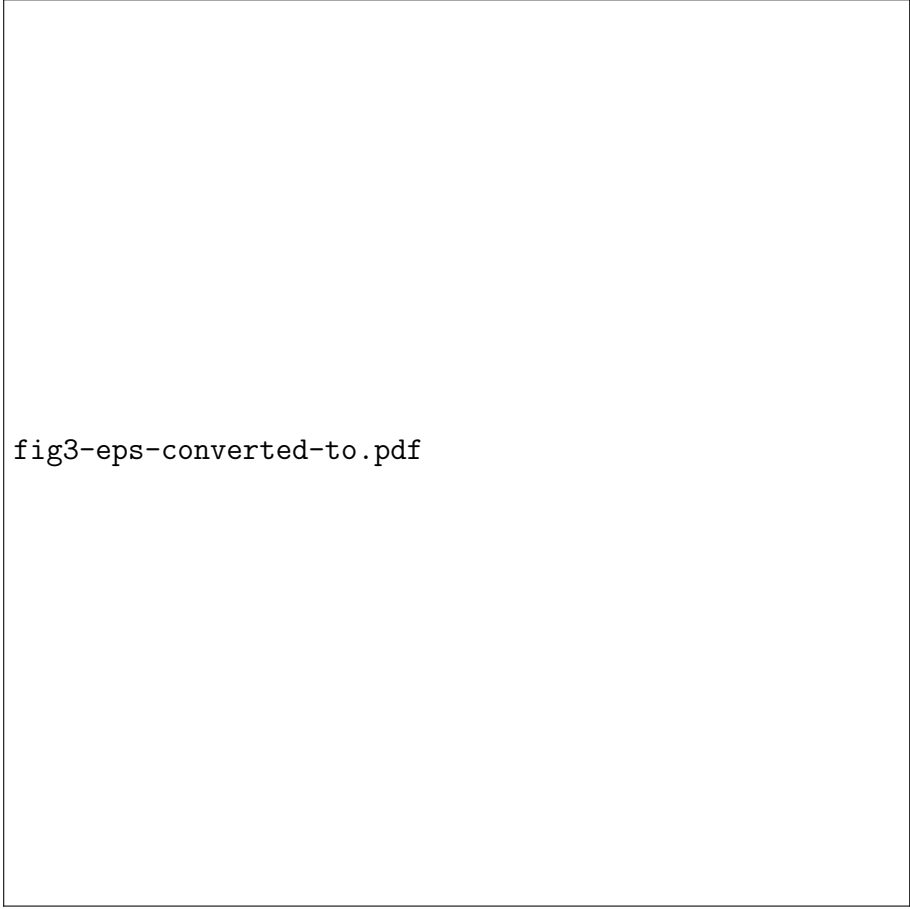


fig3-eps-converted-to.pdf

Figure 3: Relationship between u_k and u_{k+1} when $u_k = 0$

regardless of the class of customer $k + 1$. The distributions of the quantities g_{k+1} and h_{k+1} , however, do depend on the class of customer $k + 1$. More specifically, we have

$$\begin{aligned}
 G_A(z) &\triangleq E[z^{g_{k+1}} | t_{k+1} = A] = A(E(z)) \quad , \quad H_A(z) \triangleq E[z^{h_{k+1}} | t_{k+1} = A] = F(z)A(E(z)) \quad ; \\
 G_B(z) &\triangleq E[z^{g_{k+1}} | t_{k+1} = B] = B(E(z)) \quad , \quad H_B(z) \triangleq E[z^{h_{k+1}} | t_{k+1} = B] = F(z)B(E(z)) \quad .
 \end{aligned}
 \tag{8}$$

3.2 System content at customer departure times

Let us assume that the queueing system at hand is stable, i.e., that the stability condition [8, 3, 29] is fulfilled. Intuitively, it is not difficult to see that the system is stable if and only if the average amount of work entering the system per slot is strictly less than 1, i.e., if and only if

$$\lambda \cdot C'(1) < 1 \quad ,$$

where $C(z)$ denotes the pgf of the service time of an arbitrary customer given by

$$C(z) \triangleq t_A A(z) + t_B B(z) \quad , \tag{9}$$

and $C'(1)$ its average value. Expressed in the basic parameters of our system, this is equivalent to the condition

$$\lambda(t_A\mu_A + t_B\mu_B) < 1 \quad , \quad (10)$$

where, in the above expressions, the quantities t_A and t_B are the steady-state probabilities of the arrival Markov chain, defined in equation (2). Let us also denote by ρ the total load of the queueing system, defined as

$$\rho \triangleq \lambda C'(1) = \lambda[t_A\mu_A + t_B\mu_B] \quad , \quad (11)$$

and the stability condition requires that $\rho < 1$.

Assuming this condition is fulfilled, we define the joint steady-state probabilities of the Markov chain $\{(t_k, u_k)\}$ as

$$p_A(i) \triangleq \lim_{k \rightarrow \infty} \Pr[t_k = A, u_k = i] \quad , \quad p_B(i) \triangleq \lim_{k \rightarrow \infty} \Pr[t_k = B, u_k = i] \quad , \quad (12)$$

for all $i \geq 0$. The corresponding partial pgf's are given by

$$P_A(z) \triangleq \sum_{i=0}^{\infty} p_A(i)z^i \quad , \quad P_B(z) \triangleq \sum_{i=0}^{\infty} p_B(i)z^i \quad , \quad (13)$$

and the steady-state pgf of the total system content at customer departure times is equal to

$$P(z) = P_A(z) + P_B(z) \quad . \quad (14)$$

Then, if we take the z -transform of the balance equations for the Markov chain $\{(t_k, u_k)\}$ that follow from the two system equations in (6), we can establish two linear equations for the partial pgf's $P_A(z)$ and $P_B(z)$ (we refer to [9] for more details)

$$\begin{aligned} [z - \alpha A(E(z))] P_A(z) - (1 - \beta)A(E(z))P_B(z) \\ = \frac{E(z) - 1}{1 - E(0)} [\alpha P_A(0) + (1 - \beta)P_B(0)] A(E(z)) \quad , \end{aligned} \quad (15)$$

$$\begin{aligned} [z - \beta B(E(z))] P_B(z) - (1 - \alpha)B(E(z))P_A(z) \\ = \frac{E(z) - 1}{1 - E(0)} [\beta P_B(0) + (1 - \alpha)P_A(0)] B(E(z)) \quad . \end{aligned} \quad (16)$$

Explicit expressions for the two partial pgf's $P_A(z)$ and $P_B(z)$ obviously can be found by solving this set. Using these results in equation (14), we obtain the following expression for the pgf $P(z)$:

Theorem 1 *The pgf of the system content at customer departure times is given by*

$$P(z) = \frac{(1 - \rho)[E(z) - 1]}{\lambda} \frac{z[p_A A(E(z)) + p_B B(E(z))] + (1 - \alpha - \beta)A(E(z))B(E(z))}{z^2 - z[\alpha A(E(z)) + \beta B(E(z))] - (1 - \alpha - \beta)A(E(z))B(E(z))} , \quad (17)$$

The quantities p_A and p_B in the above expression denote the conditional probabilities that a customer entering an empty system (in the steady state) belongs to class A or B , respectively. As shown in [9], these quantities can be calculated from

$$p_A \triangleq \frac{\alpha P_A(0) + (1 - \beta)P_B(0)}{P_A(0) + P_B(0)} = \frac{\alpha A(E(\hat{z})) - (1 - \beta)B(E(\hat{z})) - \hat{z}}{A(E(\hat{z})) - B(E(\hat{z}))} \quad (18)$$

$$p_B \triangleq \frac{(1 - \alpha)P_A(0) + \beta P_B(0)}{P_A(0) + P_B(0)} = \frac{\beta B(E(\hat{z})) - (1 - \alpha)A(E(\hat{z})) - \hat{z}}{B(E(\hat{z})) - A(E(\hat{z}))} ,$$

where \hat{z} represents the zero inside the closed unit disk of the complex z -plane of the denominator in expression (17) for $P(z)$. Once the zero \hat{z} has been computed (numerically), p_A and p_B can be derived from (18). Substitution of the obtained values in equation (17) then leads to a fully determined expression of the steady-state pgf $P(z)$ of the total system content at customer departure times.

For further use, also note that the probability $P(0)$ is given by

$$P(0) = \frac{1 - E(0)}{\lambda} [1 - \lambda(t_A \mu_A + t_B \mu_B)] = \frac{1 - E(0)}{\lambda} (1 - \rho) . \quad (19)$$

3.3 System content at random slot boundaries

It has been shown in [8, 7] that in any discrete-time queueing system with one single server and independent arrivals from slot to slot (with pgf $E(z)$), regardless of the precise characteristics of the service process and the intra-slot details of the arrival process (the position of the arrival instants within the slot, single arrivals or batch arrivals, etc.), the following simple relationship is valid between the pgf $S(z)$ of the system content at random slot boundaries and the pgf $P(z)$ valid at customer departure times:

$$P(z) = \frac{E(z) - 1}{\lambda(z - 1)} S(z) . \quad (20)$$

Combining this general result with equation (17), we obtain the following result:

Theorem 2 *The pgf of the system content at random slot boundaries equals*

$$S(z) = (1 - \rho)(z - 1) \frac{z[p_A A(E(z)) + p_B B(E(z))] + (1 - \alpha - \beta)A(E(z))B(E(z))}{z^2 - z[\alpha A(E(z)) + \beta B(E(z))] - (1 - \alpha - \beta)A(E(z))B(E(z))} . \quad (21)$$

From this expression, various performance measures of practical importance can be derived. For instance, the mean system content at random slot marks can be found as $E[s] = S'(1)$. After long and tedious calculations, we find

Theorem 3 *The mean system content at customer departure times is given by*

$$E[s] = \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} + \frac{\gamma t_A t_B \lambda^2 (\mu_A - \mu_B)^2}{(1-\gamma)(1-\rho)} + \frac{\lambda(p_A - t_A)(\mu_A - \mu_B)}{1-\gamma}. \quad (22)$$

In the above expression, t_A and t_B are given in equation (2), γ is the interclass correlation defined in (3), $C'(1)$ and $C''(1)$ are derivatives of the pgf $C(z)$ of the service time of an arbitrary customer, ρ is the load of the system calculated from (11), and p_A and p_B are the probabilities given by the formulas (18).

The first term (ρ) in equation (22) corresponds to the mean number of customers in service, the other three terms account for the mean *queue content*, i.e., the mean number of customers that are actually waiting to be served.

Higher-order moments of the system-content distribution can be obtained by computing higher-order derivatives of the pgf $S(z)$. By applying (the discrete-time version of) Little's law [29, 8, 20], the mean *delay* (system time) of an arbitrary customer can be obtained as $E[d] = E[s]/\lambda$. The mean *waiting time* of an arbitrary customer can be derived from this as $E[w] = E[d] - C'(1)$ and is given by

$$E[w] = \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2\lambda(1-\rho)} + \frac{\gamma t_A t_B \lambda (\mu_A - \mu_B)^2}{(1-\gamma)(1-\rho)} + \frac{(p_A - t_A)(\mu_A - \mu_B)}{1-\gamma}. \quad (23)$$

4 Customer waiting time analysis

Whereas a formula for the mean customer waiting time was given in the previous section, in this section we derive the steady-state pgf of this random variable. As we will see in the next subsection, this also leads to some new and interesting results on the joint pgf of the system content at random slot boundaries and the residual service time as well.

4.1 System content and residual service time

In order to be able to do waiting time calculations, we first need to consider the system content at random slot boundaries, and the residual service time of the customer being served (if any). We do this step-by-step. To set our minds, let us first take a look at the system with only one class of customers, for instance by setting $A(z) = B(z) = C(z)$ (see Section 5 on some results for the single-class system), implying that all customers have the same service time distribution.

Single – class system

Consider an arbitrarily selected customer \tilde{C} and denote its arrival slot by \tilde{R} ; hence slot \tilde{R} is a randomly selected slot as well. Let us also define \tilde{u} as the number of customers in the system just after the service completion that precedes slot \tilde{R} . Note that u and \tilde{u} do not have the same

distribution, since $\tilde{u} = 0$ implies that \tilde{R} either falls within an idle period, or the first customer service time of a busy period. However, if we denote, as before, by s (with pgf $S(z)$) the system content at the beginning of a randomly chosen slot \tilde{R} , then conditioning on $s > 0$ implies that \tilde{R} falls within a (random) customer service time, and choosing a random slot then becomes equivalent to choosing a random service completion. Therefore,

$$\Pr[\tilde{u} = j | s > 0] = \Pr[u = j] \quad , \quad j \geq 0 \quad . \quad (24)$$

In order to confirm this result, consider $j = 0$, and let $E[Id]$ and $E[Bu]$ represent the average length of an idle and busy period respectively. For the system under consideration, these quantities satisfy (see [8])

$$E[Id] = \frac{1}{1 - E(0)} \quad ; \quad E[Bu] = \frac{\rho}{1 - \rho} \cdot \frac{1}{1 - E(0)} \quad .$$

Note that $\Pr[\tilde{u} = 0, s > 0]$ is equal to the probability that slot \tilde{R} falls within the first customer service time of a busy period, and hence, with $\Pr[s = 0] = S(0) = 1 - \rho$,

$$\Pr[\tilde{u} = 0 | s > 0] = \rho^{-1} \frac{\mu_C}{E[Id] + E[Bu]} = \frac{1 - \rho}{\lambda} \cdot (1 - E(0)) \quad , \quad (25)$$

with μ_C the mean length of a (random) service time. This is indeed equal to $\Pr[u = 0] = P(0)$ (see (29) further on).

Throughout this section, $s > 0$ represents the event where slot \tilde{R} falls within a customer service time, and let us define c_a (c_b), with pgf $C_a(z)$ ($C_b(z)$) as the residual (elapsed) customer service time that follows (precedes) slot \tilde{R} (by convention, we assume that slot \tilde{R} is included in c_a , but not in c_b). For $s = 0$ we set $c_a = c_b = 0$, and in case of $s > 0$, a simple extension of the standard residual waiting time calculation reveals that the joint pgf of these two random variables satisfies

$$E[x^{c_a} y^{c_b}] = x \frac{C(x) - C(y)}{\mu_C(x - y)} \quad . \quad (26)$$

By setting $y = 1$ ($x = 1$) in this expression, we obtain $C_a(x)$ ($C_b(y)$). Some relevant quantities that appear in the subsequent analysis are depicted in Fig. 4.

For $s > 0$, we can now express s in terms of \tilde{u} by means of the system equations

$$s = \begin{cases} f + 1 + \sum_{i=1}^{c_b} e_i \quad , & \text{if } \tilde{u} = 0 \\ \tilde{u} + \sum_{i=1}^{c_b} e_i \quad , & \text{if } \tilde{u} > 0 \quad , \end{cases} \quad (27)$$

where f (with pgf $F(z)$; see equation (7)) represents the number of additional arrivals in slot with at least one arrival, and with e_i the number of customer arrivals during the i -th slot of c_b . If we translate these system equations into z -transforms, then by invoking (24) we find the following relation

$$S(z) = 1 - \rho + \rho C_b(E(z)) \cdot \{P(z) + P(0)(zF(z) - 1)\} \quad . \quad (28)$$

fig4-eps-converted-to.pdf

Figure 4: arrival of \tilde{C} in case $s > 0$: definitions

With $P(z)$ satisfying (see [17])

$$P(z) = \frac{1 - \rho}{\lambda} \cdot \frac{C(E(z))(E(z) - 1)}{z - C(E(z))} \quad , \quad (29)$$

and $S(z)$ given by (46), it is easily verified that (28) indeed holds.

As a next step, let us consider the couple (s, c_a) (we still focus on the system with one customer class). We define its joint pgf as

$$S_a(z, x) \triangleq E[z^s x^{c_a}] \quad .$$

Since \tilde{u} refers to the service completion that precedes slot \tilde{R} , this random variable is independent from c_a and c_b . Therefore, invoking system equation (27), and making use of (24) and (26) while following the same approach as above, we find that

$$S_a(z, x) = 1 - \rho + \rho x \frac{C(x) - C(E(z))}{C'(1)(x - E(z))} \cdot \{P(z) + P(0)(zF(z) - 1)\} \quad .$$

With expression (29) for $P(z)$, this becomes

$$S_a(z, x) = (1 - \rho) \left(1 + xz \frac{C(x) - C(E(z))}{x - E(z)} \cdot \frac{E(z) - 1}{z - C(E(z))} \right) .$$

The same result, by means of a direct analysis of the two-dimensional state description (s, c_a) , was also obtained in [8].

Two – class system

Let us now focus on the system with two classes of customers, i.e. $A(z) \neq B(z)$, and where the class of consecutive customers is designated by the two-state Markov process described in Section 2. In addition to the previous definitions, for $s > 0$, let us now define \tilde{t} as the class of customer being served during slot \tilde{R} . We first establish some preliminary results that will be used later on. Since slot \tilde{R} is selected by a random customer arrival, $\Pr[\tilde{t} = A | s > 0]$ is proportional to the relative frequency of class- A customers in the arrival process, as well as to their average service time. Hence we can write, relying on a similar reasoning as in (24) and with $\mu_C \triangleq t_A \mu_A + t_B \mu_B = \rho / \lambda$

$$\Pr[\tilde{u} = j, \tilde{t} = A | s > 0] = \frac{\mu_A}{\mu_C} \hat{p}_A(j) ; \Pr[\tilde{u} = j, \tilde{t} = B | s > 0] = \frac{\mu_B}{\mu_C} \hat{p}_B(j) ,$$

with

$$\begin{aligned} \hat{p}_A(j) &\triangleq \lim_{k \rightarrow \infty} \Pr[u_k = j, t_{k+1} = A] = \alpha p_A(j) + (1 - \beta) p_B(j) \\ \hat{p}_B(j) &\triangleq \lim_{k \rightarrow \infty} \Pr[u_k = j, t_{k+1} = B] = (1 - \alpha) p_A(j) + \beta p_B(j) , \end{aligned}$$

and with corresponding partial pgfs $\hat{P}_A(z)$ and $\hat{P}_B(z)$ respectively. In addition, note that from (15), it follows that

$$\begin{aligned} \hat{P}_A(z) + \hat{P}_A(0)(zF(z) - 1) &= z \frac{P_A(z)}{A(E(z))} \\ \hat{P}_B(z) + \hat{P}_B(0)(zF(z) - 1) &= z \frac{P_B(z)}{B(E(z))} . \end{aligned} \quad (30)$$

Let us define the partial joint pgfs

$$\tilde{S}_A(z, x) \triangleq E[z^s x^{c_a} \mathbf{I}_{\{\tilde{t}=A\}} | s > 0] ; \tilde{S}_B(z, x) \triangleq E[z^s x^{c_a} \mathbf{I}_{\{\tilde{t}=B\}} | s > 0] ,$$

with $\mathbf{I}_{\{X\}}$ the indicator function, which equals 1 if the event X is true, and 0 otherwise. If $\tilde{t} = A$ (resp B), then c_a and c_b obviously represent the class- A (resp. B) residual customer service times after and before slot \tilde{R} , and similar expressions as in (26) must be adopted accordingly in the analysis. Therefore, if we again invoke system equation (27), which is still valid for the 2-class system, and further consider the two possible values of \tilde{t} , then with the above results, we obtain the following expressions for these partial joint pgfs:

$$\tilde{S}_A(z, x) = xz \frac{P_A(z)}{A(E(z))} \frac{A(x) - A(E(z))}{\mu_C(x - E(z))}$$

$$\tilde{S}_B(z, x) = xz \frac{P_B(z)}{B(E(z))} \frac{B(x) - B(E(z))}{\mu_C(x - E(z))} . \quad (31)$$

Thus, finally, the joint pgf of the couple (s, c_a) is given by

$$S_a(z, x) = 1 - \rho + \rho(\tilde{S}_A(z, x) + \tilde{S}_B(z, x)) . \quad (32)$$

Remark

Let us verify these results by means of a small additional calculation. From (31) and (32), and with $P(z) = P_A(z) + P_B(z)$, we find for $S(z) \triangleq S_a(z, 1)$

$$\frac{S(z)}{\lambda} = \frac{(z-1)P(z)}{E(z)-1} + \frac{1-\rho}{\lambda} - \frac{P_A(z)}{A(E(z))} \frac{z-A(E(z))}{E(z)-1} - \frac{P_B(z)}{B(E(z))} \frac{z-B(E(z))}{E(z)-1} .$$

On the other hand, note that if we divide (15) by $A(E(z))$ and (16) by $B(E(z))$, respectively, then the sum of the two equations that we thus obtain yields

$$\frac{P_A(z)}{A(E(z))}(z - A(E(z))) + \frac{P_B(z)}{B(E(z))}(z - B(E(z))) = \frac{E(z) - 1}{E(0) - 1} P(0) = \frac{1 - \lambda}{\rho} (E(z) - 1) ,$$

where we have invoked (25), which is still valid for the 2-class system. Hence, from the two above equations, we find the same relation between $P(z)$ and $S(z)$ as in (20), thereby confirming our derivations.

In our calculations of the next section, for $s > 0$, we require the joint partial pgf of $(q, c_a - 1)$, where q represents the queue content at the beginning of slot \tilde{R} , i.e., the number of customers in the system, not including the one being served. Obviously, $q = s - 1$ for $s > 0$, and we find in view of the previous results

$$\begin{aligned} \tilde{Q}(z, x) &\triangleq E[z^q x^{c_a-1} | s > 0] \\ &= \frac{P_A(z)}{A(E(z))} \frac{A(x) - A(E(z))}{\mu_C(x - E(z))} + \frac{P_B(z)}{B(E(z))} \frac{B(x) - B(E(z))}{\mu_C(x - E(z))} . \end{aligned} \quad (33)$$

4.2 Customer waiting time

We define the random variable w , with steady-state pgf $W(z)$, as the waiting time of customer \tilde{C} , which is equal to the number of slots between the end of slot \tilde{R} , and the beginning of the slot during which \tilde{C} 's service time is initiated. Consider the case that $s > 0$. Under a FCFS discipline, w is determined by the class of the customer being served during slot \tilde{R} , its residual service time, and the number of class- A and B customers found in the queue by \tilde{C} upon its arrival. In order to determine this number of class- A and B customers, we need to take into account the specifics of the 2-state Markov process that designates the class of consecutive customers. Given that a class- A customer is being served during slot \tilde{R} and that \tilde{C} finds n customers in the queue upon arrival, let $t_{A,n}$ (with pgf $T_{A,n}(z)$) denote the number of class- A customers in the queue; similarly, we let $t_{B,n}$ (with pgf $T_{B,n}(z)$) represent the number of class- B customers in the queue in case that a class- B customer is being served. Obviously, $n - t_{A,n}$

and $n - t_{B,n}$ then represents the number of class- B customers in the queue in both cases. With $T_{A,0}(z) = T_{B,0}(z) = 1$, it is easy to check that, for $n \geq 1$, the conditional pgf's $T_{A,n}(z)$ and $T_{B,n}(z)$ can be calculated by the recursive matrix equation

$$\begin{aligned} \begin{bmatrix} T_{A,n}(z) \\ T_{B,n}(z) \end{bmatrix} &= \begin{bmatrix} \alpha z & (1 - \alpha) \\ (1 - \beta)z & \beta \end{bmatrix} \begin{bmatrix} T_{A,n-1}(z) \\ T_{B,n-1}(z) \end{bmatrix} \\ &= \begin{bmatrix} \alpha z & (1 - \alpha) \\ (1 - \beta)z & \beta \end{bmatrix}^n \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned} \quad (34)$$

This indicates that the matrix $\mathbf{Q}(z)$, defined as

$$\mathbf{Q}(z) \triangleq \begin{bmatrix} \alpha z & (1 - \alpha) \\ (1 - \beta)z & \beta \end{bmatrix},$$

and its diagonalized form, will play an important role in the subsequent analysis. The eigenvalues of $\mathbf{Q}(z)$ are given by

$$\begin{aligned} \lambda_1(z) &= \frac{1}{2} \left[\alpha z + \beta + \sqrt{(\alpha z - \beta)^2 + 4(1 - \alpha)(1 - \beta)z} \right] \\ \lambda_2(z) &= \frac{1}{2} \left[\alpha z + \beta - \sqrt{(\alpha z - \beta)^2 + 4(1 - \alpha)(1 - \beta)z} \right]. \end{aligned} \quad (35)$$

In addition, the 2×2 matrix $\mathbf{R}(z)$ that contains the right-column eigenvectors of $\mathbf{Q}(z)$ satisfies

$$\mathbf{R}(z) = \begin{bmatrix} \frac{1 - \alpha}{\lambda_1(z) - \alpha z} \frac{\lambda_1(z) + (1 - \alpha - \beta)z}{\lambda_1(z) - \lambda_2(z)} & \frac{1 - \alpha}{\lambda_2(z) - \alpha z} \frac{\lambda_2(z) + (1 - \alpha - \beta)z}{\lambda_2(z) - \lambda_1(z)} \\ \frac{\lambda_1(z) + (1 - \alpha - \beta)z}{\lambda_1(z) - \lambda_2(z)} & \frac{\lambda_2(z) + (1 - \alpha - \beta)z}{\lambda_2(z) - \lambda_1(z)} \end{bmatrix},$$

where we have relied on the normalization condition

$$\mathbf{R}(z) \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbf{R}(z)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (36)$$

Diagonalization of $\mathbf{Q}(z)$ in (34) then leads to, in view of (36)

$$\begin{bmatrix} T_{A,n}(z) \\ T_{B,n}(z) \end{bmatrix} = \mathbf{R}(z) \begin{bmatrix} \lambda_1(z)^n \\ \lambda_2(z)^n \end{bmatrix}. \quad (37)$$

Let us briefly point to some of the properties of these conditional pgfs. From the above expressions, it follows that the components of $\mathbf{R}(z)$ have both a simple pole and a branch point for values of $z = \hat{z}$ that satisfy $\lambda_1(\hat{z}) = \lambda_2(\hat{z})$ (i.e., the square root that appears in (35) equals 0 for $z = \hat{z}$, implying that \hat{z} is the solution of a quadratic equation), and some \hat{z} may even fall within the complex unit circle $|z| < 1$. However, a closer inspection reveals that the \hat{z} are *removable singular points* ([24]) of the conditional pgfs in (37). Let us for instance take a look at

$$T_{B,n}(z) = \frac{\lambda_1(z) + (1 - \alpha - \beta)z}{\lambda_1(z) - \lambda_2(z)} \lambda_1(z)^n + \frac{\lambda_2(z) + (1 - \alpha - \beta)z}{\lambda_2(z) - \lambda_1(z)} \lambda_2(z)^n.$$

Consider any function $f(z)$ for which $f(z) \in \mathbb{R}$, $\forall z \in \mathbb{R}$. Then the MacLaurin series expansion of $f(x + y) + f(x - y)$ around $y = 0$ is an even function of y , i.e., contains only even powers

of y . If we let y represent the square root that appears in (35), then this reasoning shows that $T_{B,n}(z)$ is an even function of y and therefore contains neither a branch point nor a simple pole for $y = 0$ ¹. Indeed, if we work out the above expression for $T_{B,n}(z)$ in terms of (even) powers of the square root in (35), we obtain

$$T_{B,n}(z) = \left(\frac{1}{2}\right)^n (\beta + (\alpha + 2(1 - \beta))z) \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n}{2j+1} (\alpha z + \beta)^{n-2j-1} \\ ((\alpha z - \beta)^2 + 4(1 - \alpha)(1 - \beta)z)^j \\ + \left(\frac{1}{2}\right)^n \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2j} (\alpha z + \beta)^{n-2j} ((\alpha z - \beta)^2 + 4(1 - \alpha)(1 - \beta)z)^j ,$$

with the convention that the first sum in the right-hand side equals 0 for $n = 0$. This expression also shows that $T_{B,n}(z)$ is a polynomial of degree n , as expected. A similar expression can be established for $T_{A,n}(z)$.

Next, let us return to the waiting time calculation, still under the assumption that $s > 0$, and define θ as the residual burst size, being the number of customers that arrive during slot \tilde{R} and will be served before customer \tilde{C} (see Fig. 4). Its pgf $\Theta(z)$ satisfies ([8])

$$\Theta(z) \triangleq E[z^\theta] = \frac{E(z) - 1}{\lambda(z - 1)} .$$

Then $q + \theta$ is the number of customers that will receive full service during the waiting time of \tilde{C} . Relying on the above calculations, we can determine how many of these customers are of class A and B , and we let a_j (b_j), with pgf $A(z)$ ($B(z)$) as before, represent the respective service times of the j -th class- A (B) customer in the queue. In addition, we must take into account the residual service time c_a of the customer being served during slot \tilde{R} . Since c_a includes slot \tilde{R} while w does not, we can express the conditional pgf of w as

$$E[z^w | c_a - 1 = i, q + \theta = n, \tilde{t} = A, s > 0] \\ = E \left[z^{i + \sum_{j=1}^{t_{A,n}} a_j + \sum_{j=1}^{n-t_{A,n}} b_j} \middle| c_a - 1 = i, q + \theta = n, \tilde{t} = A, s > 0 \right] = z^i B(z)^n T_{A,n}(\phi(z)) ,$$

for $i, n \geq 0$, and with $\phi(z) \triangleq \frac{A(z)}{B(z)}$. An analogous expression evidently holds in case of $\tilde{t} = B$. The random variable θ and the triplet (q, c_a, \tilde{t}) are mutually independent. If we take into account expression (37) for $T_{A,n}(z)$ and $T_{B,n}(z)$ and average the above expected value over all possible values of $c_a - 1$, q , θ and \tilde{t} , and keep in mind our definition of $\tilde{Q}(z, x)$ in (33), we eventually obtain

$$E[z^w | s > 0] = [1 \quad 1] \mathbf{R}(\phi(z)) \begin{bmatrix} \Theta(B(z)\lambda_1(\phi(z)))\tilde{Q}(B(z)\lambda_1(\phi(z)), z) \\ \Theta(B(z)\lambda_2(\phi(z)))\tilde{Q}(B(z)\lambda_2(\phi(z)), z) \end{bmatrix} . \quad (38)$$

¹Notice the analogy with complex numbers calculus: for $z = x + iy$ with $x, y \in \mathbb{R}$, then $f(z) + f(z^*) \in \mathbb{R}$, since this is an even function of iy

Finally, in case of $s = 0$, slot \tilde{R} falls within an idle period, and define \tilde{t}_p as the class of the last customer that was served before slot \tilde{R} . Due to the i.i.d. nature of the aggregate arrival process, all idle periods have the same (geometric) distribution, regardless of the last customer's class. Therefore, since $P_A(0)$ ($P_B(0)$) equals the probability that a class- A (B) customer leaves behind an empty system, we have that

$$\tilde{q}_A \triangleq \Pr[\tilde{t}_p = A] = \frac{P_A(0)}{P_A(0) + P_B(0)} \quad ; \quad \tilde{q}_B \triangleq \Pr[\tilde{t}_p = B] = \frac{P_B(0)}{P_A(0) + P_B(0)}$$

The waiting time of \tilde{C} is now determined by the value of θ in the following way:

$$\begin{aligned} E[z^w | \theta = n, \tilde{t}_p = A, s = 0] &= B(z)^n T_{A,n}(\phi(z)) \\ E[z^w | \theta = n, \tilde{t}_p = B, s = 0] &= B(z)^n T_{B,n}(\phi(z)) . \end{aligned}$$

If we average these expected values over all possible values of θ and \tilde{t}_p , we thus find

$$E[z^w | s = 0] = [\tilde{q}_A \quad \tilde{q}_B] \mathbf{R}(\phi(z)) \begin{bmatrix} \Theta(B(z)\lambda_1(\phi(z))) \\ \Theta(B(z)\lambda_2(\phi(z))) \end{bmatrix} \quad (39)$$

Finally, with (38)-(39), we obtain:

Theorem 4 *The pgf of the customer waiting time can be expressed as*

$$\begin{aligned} W(z) &= (1 - \rho) [\tilde{q}_A \quad \tilde{q}_B] \mathbf{R}(\phi(z)) \begin{bmatrix} \Theta(B(z)\lambda_1(\phi(z))) \\ \Theta(B(z)\lambda_2(\phi(z))) \end{bmatrix} \\ &+ \rho \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{R}(\phi(z)) \begin{bmatrix} \Theta(B(z)\lambda_1(\phi(z)))\tilde{Q}(B(z)\lambda_1(\phi(z)), z) \\ \Theta(B(z)\lambda_2(\phi(z)))\tilde{Q}(B(z)\lambda_2(\phi(z)), z) \end{bmatrix} . \end{aligned} \quad (40)$$

Note that, due to $\tilde{q}_A + \tilde{q}_B = 1$, $B(1) = \phi(1) = \lambda_1(1) = \Theta(1) = \tilde{Q}(1) = 1$, and $\mathbf{R}(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, one can easily verify that the right-hand side of this expression equals 1 for $z = 1$, implying that $W(1) = 1$, i.e., this pgf is indeed normalized.

5 Three special cases

5.1 Identical service-time distributions: $A(z) = B(z)$

First, let us consider the special case whereby the pgf's $A(z)$ and $B(z)$ are identical. The pgf $S(z)$ can then be rewritten as

$$\begin{aligned} S(z) &= (1 - \rho)(z - 1) \frac{zA(E(z)) + (1 - \alpha - \beta)A^2(E(z))}{z^2 - (\alpha + \beta)zA(E(z)) - (1 - \alpha - \beta)A^2(E(z))} \\ &= (1 - \rho)(z - 1) \frac{A(E(z))[z + (1 - \alpha - \beta)A(E(z))]}{[z - A(E(z))][z + (1 - \alpha - \beta)A(E(z))]} \\ &= (1 - \rho) \frac{(z - 1)A(E(z))}{z - A(E(z))} . \end{aligned} \quad (41)$$

This is the well-known [8, 7] pgf of the system content at random slot boundaries in a single-server system with one class of customers with service-time pgf $A(z)$, as expected.

Let us also consider the result for $W(z)$. For $A(z) = B(z)$, we have that $\phi(z) = 1$, and therefore also $\lambda_1(\phi(z)) = 1$ and $\mathbf{R}(\phi(z)) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. Then, in view of the definition of $\tilde{Q}(z, x)$, the results of the previous section immediately lead to

$$W(z) = \Theta(A(z)) \left(1 - \rho + \frac{1}{zA(z)} (S_a(A(z), z) - (1 - \rho)) \right) . \quad (42)$$

If we now invoke the explicit expressions for $\Theta(z)$ and $S_a(z, x)$, we eventually find

$$W(z) = \frac{1 - \rho}{\lambda} \frac{E(A(z)) - 1}{A(z) - 1} \frac{z - 1}{z - E(A(z))} . \quad (43)$$

An expression for the steady-state pgf of the system time (i.e., waiting time + service time) for the single-class case can be found in [8, 7] as well, and is indeed equal to the above expression for $W(z)$ multiplied by $A(z)$.

5.2 No interclass correlation: $\gamma = 0$

A second remarkable special case is obtained when the interclass correlation γ is zero, i.e., when $\alpha + \beta = 1$ or $\beta = 1 - \alpha$. In this case, the two customer classes occur randomly and independently in the arrival stream and the distribution of all service times is given by

$$C(z) \triangleq \alpha A(z) + (1 - \alpha)B(z) , \quad (44)$$

independently from customer to customer.

Formula (21) then reduces to

$$S(z) = (1 - \rho)(z - 1) \frac{z[p_A A(E(z)) + p_B B(E(z))]}{z^2 - z[\alpha A(E(z)) + (1 - \alpha)B(E(z))]} .$$

In this case, the quantity \hat{z} is equal to zero, and equations (18) reduce to $p_A = \alpha$ and $p_B = 1 - \alpha$. In these circumstances, $S(z)$ can be rewritten as

$$\begin{aligned} S(z) &= (1 - \rho)(z - 1) \frac{\alpha A(E(z)) + (1 - \alpha)B(E(z))}{z - [\alpha A(E(z)) + (1 - \alpha)B(E(z))]} \\ &= (1 - \rho) \frac{(z - 1)C(E(z))}{z - C(E(z))} . \end{aligned} \quad (45)$$

Again, we obtain the well-known [8, 7] steady-state pgf of the system content at random slot boundaries in a single-server system with one class of customers with service-time pgf $C(z)$, as expected.

As far as the calculations for $W(z)$ are concerned, we now have that $\lambda_1(z) = 1 - \alpha + \alpha z$, $\lambda_2(z) = 0$, and $\mathbf{R}(\phi(z)) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. Therefore, $B(z)\lambda_1(\phi(z))$ is equal to $C(z)$ given above, and the remaining calculations are therefore essentially the same as in the previous case.

5.3 Single-class system: $\alpha = 1$

As a third special case, let us assume $\alpha = 1$ and β arbitrary. In this case, only customers of class A arrive in the steady state and we expect to retrieve result (41) again. This is, indeed, the case. Formula (21) now reduces to

$$\begin{aligned} S(z) &= (1 - \rho)(z - 1) \frac{z[p_A A(E(z)) + p_B B(E(z))] - \beta A(E(z))B(E(z))}{z^2 - z[A(E(z)) + \beta B(E(z))] + \beta A(E(z))B(E(z))} \\ &= (1 - \rho)(z - 1) \frac{z[p_A A(E(z)) + p_B B(E(z))] - \beta A(E(z))B(E(z))}{[z - A(E(z))][z - \beta B(E(z))]} . \end{aligned}$$

The zero \hat{z} is a zero of the second factor in the denominator of the above expression, which implies that

$$\hat{z} = \beta B(E(\hat{z})) ,$$

so that equations (18) reduce to $p_A = 1$ and $p_B = 0$. The pgf $S(z)$ can therefore be simplified as

$$\begin{aligned} S(z) &= (1 - \rho)(z - 1) \frac{zA(E(z)) - \beta A(E(z))B(E(z))}{[z - A(E(z))][z - \beta B(E(z))]} \\ &= (1 - \rho) \frac{(z - 1)A(E(z))}{z - A(E(z))} , \end{aligned} \tag{46}$$

as expected.

For the calculations for $W(z)$ we find that $\lambda_1(z) = z$, $\lambda_2(z) = \beta z$, and $\mathbf{R}(\phi(z)) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. Therefore, $B(z)\lambda_1(\phi(z))$ is equal to $A(z)$, and again, we end up with similar calculations as before, to yield (43).

6 Discussion of results and numerical examples

In this section, we discuss the results obtained for the general case, both from a qualitative perspective and by means of some numerical examples.

The first interesting result obtained is the form of the stability condition (10),

$$\lambda < \frac{1}{C'(1)} = \frac{1}{t_A \mu_A + t_B \mu_B} ,$$

which shows that the maximum achievable throughput of this system, expressed in customers per slot, is completely determined by the mean service time of an arbitrary customer, regardless of the possible interclass correlation.

Next, for sake of brevity, our main focus in this and the next section will be on the formula (22) that we obtained for the mean system content at random slot marks:

$$E[s] = \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} + \frac{\gamma t_A t_B \lambda^2 (\mu_A - \mu_B)^2}{(1-\gamma)(1-\rho)} + \frac{\lambda(p_A - t_A)(\mu_A - \mu_B)}{1-\gamma} . \quad (47)$$

This result very clearly shows the influence of the various system parameters on the performance of the system. As could be expected intuitively, the mean system content depends on the first two moments of the arrival process (as represented by the quantities λ and $E''(1)$), and to some extent $\rho = \lambda C'(1)$ and the first two moments of the service times (contained in the quantities $C'(1)$, $C''(1)$, μ_A , μ_B , and also $\rho = \lambda C'(1)$). It is not surprising that $E[s]$ goes to infinity as ρ approaches its limiting value 1, dictated by the stability condition of the system. However, it is striking that $E[s]$ also seems to increase without bound if the interclass correlation $\gamma = \alpha + \beta - 1$ approaches the value +1, even when the stability condition $\rho < 1$ is met. Positive interclass correlation appears to be very detrimental for the performance of the system, whereas negative interclass correlation has a very moderate positive effect on the performance. In fact, the truth is a little more subtle, as will be discussed in section 7.

The first two terms in formula (47) correspond to the classical result that would be obtained in a system without interclass correlation and with service-time pgf $C(z)$ (see e.g. [8, 7]). This means that the third and fourth terms in (47) can be fully attributed to the presence of class clustering in the arrival process. We note, indeed, that the third term vanishes when $\gamma = 0$; in the fourth term, both t_A and p_A reduce to the same value α when $\gamma = 0$ (see equations (2) and (18) with $\hat{z} = 0$), which implies that the fourth term is equal to zero as well in that case. It is easy to see that the third and fourth terms also disappear when all customers have the same service-time distribution, i.e., when $A(z) = B(z)$ and, hence $\mu_A = \mu_B$, and, finally, also in the case where there is only one class of customers in the system, i.e., where either $\alpha = 1$ (and, hence, $p_A = t_A = 1$ and $t_B = 0$) or $\beta = 1$ (and, therefore, $p_A = t_A = 0$).

Let us consider some numerical results. In a first example, we assume Poisson arrivals (i.e., $E(z) = e^{\lambda(z-1)}$), equal fractions of both classes of customers in the arrival stream (i.e., $t_A = t_B = 0.5$), geometrically distributed service times of both classes, i.e.,

$$A(z) = \frac{z}{\mu_A + (1 - \mu_A)z} \quad ; \quad B(z) = \frac{z}{\mu_B + (1 - \mu_B)z} , \quad (48)$$

with $\mu_A = 8$ and $\mu_B = 2$. The stability condition (10) is then given by $\rho = \lambda[t_A \mu_A + t_B \mu_B] = 5\lambda < 1$ (i.e., $\lambda < 0.2$).

Fig. 5 shows the mean system content $E[s]$ as a function of the load ρ for various values of the interclass correlation γ . The figure confirms that, for given values of $\rho < 1$, the parameter γ has a major impact on the results when it is positive and only a minor influence when it is negative. An intuitive explanation of this phenomenon lies in the observation that the numbers of consecutive class-A customers and class-B customers in the arrival stream both increase dramatically as γ approaches the value +1. Indeed, the mean number of class-A customers between two consecutive class-B customers is given by

$$\text{mean class-A sequence} = \frac{1}{1-\alpha} = \frac{1}{t_B(1-\gamma)} = \frac{2}{1-\gamma} ,$$

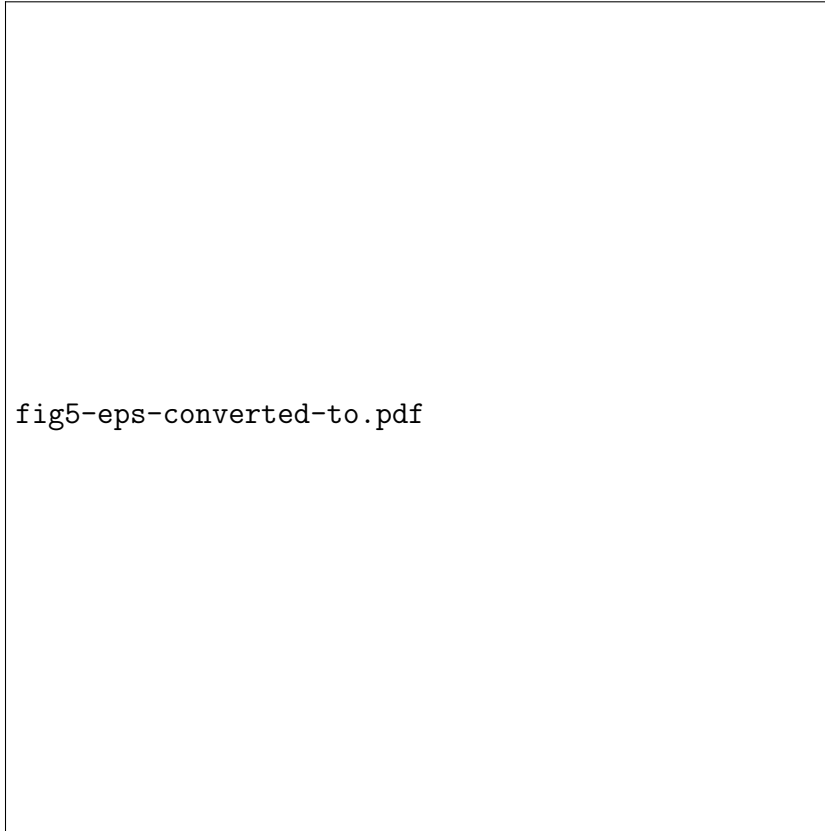


Figure 5: Mean system content $E[s]$ versus ρ , for various values of γ

and, similarly, the mean number of class-B customers between two consecutive class-A customers is equal to

$$\text{mean class-B sequence} = \frac{1}{1-\beta} = \frac{1}{t_A(1-\gamma)} = \frac{2}{1-\gamma} .$$

For negative values of γ , this implies that customers of both classes alternate strongly. For positive values of γ , however, there may be very long sequences of customers of the same class. During such periods, the momentary load is either given by

$$\rho_A \triangleq \lambda\mu_A = 8\lambda \tag{49}$$

or by

$$\rho_B \triangleq \lambda\mu_B = 2\lambda , \tag{50}$$

respectively. It is easily seen that the stability condition $\rho < 1$ or $\lambda < 0.2$ guarantees that $\rho_B = 2\lambda$ is strictly less than 1, but not necessarily that $\rho_A = 8\lambda < 1$. It is clear that, if λ or ρ is small enough (more specifically, $\lambda < 0.125$ or $\rho < 0.625$) the system is locally stable both during A-sequences and B-sequences (and, hence, also globally stable), while if $0.125 \leq \lambda < 0.2$ or, equivalently, $0.625 \leq \rho < 1$, the system is locally stable during B-sequences but not during A-sequences. In the latter case, (global) stability is assured because although the queue size builds up during A-sequences (because, on average, more work arrives than the server can

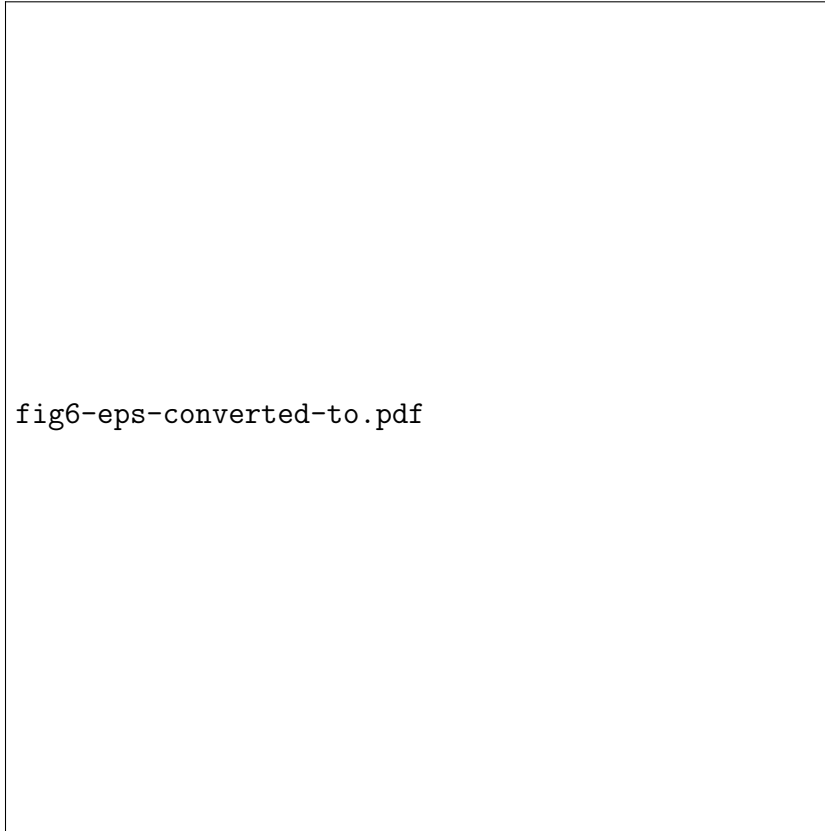


Figure 6: Mean system content $E[s]$ versus γ , for various values of ρ

perform), it decreases again during B-sequences (when much less work enters than the server can execute). However, when the interclass correlation approaches $+1$, the amplitude of these queue size variations goes to infinity, implying that the mean system content does the same.

The same behavior can be observed in Fig. 6, where we have plotted $E[s]$ as a function of γ for various values of ρ . The figure illustrates very clearly that the system content grows without bound as $\gamma \rightarrow +1$ when ρ is higher than its critical value 0.625 . However, when ρ is less than this critical value, the mean system content remains finite for all values of γ . This result, obtained numerically here and also explained intuitively, is somewhat unexpected in view of formula (47), which seems to say that the mean system content $E[s]$ should become unbounded as $\gamma \rightarrow +1$ *regardless of the other system parameters*. In order to clarify this seeming contradiction between the formula and the numerical results, we take a closer look at the limit case $\gamma \rightarrow +1$ in the next section.

A second example is treated in Figs. 7 to 9. Here, again, we assume Poisson arrivals and geometrically distributed service times of both classes (as defined in equation (48)), in this case with mean values $\mu_A = 100$ and $\mu_B = 10$. The interclass correlation γ is kept constant at $\gamma = 0.8$. This implies that $\alpha = 0.8 + 0.2t_A$, $\beta = 1 - 0.2t_A$ and $\rho = 10\lambda(1 + 9t_A)$. We now investigate the impact of the parameter t_A , i.e., the fraction of class-A customers in the arrival stream, on the mean system content and the mean waiting times of the customers. Figs. 7 and 8 show the mean system content $E[s]$ versus ρ (for various values of t_A) and versus t_A (for various values of ρ), respectively. Similarly, Fig. 9 shows the corresponding result as in Fig.



Figure 7: Mean system content $E[s]$ versus ρ , for $\gamma = 0.8$ and various values of t_A

8 for the mean waiting time $E[w]$ of the customers. Fig. 7 shows that for any given value of t_A , the mean system content is an increasing function of the load ρ , as expected. Furthermore, Fig. 8 reveals that, for any given value of the total load ρ , the mean system content increases as a function of t_A for “low” values of t_A (more or less in the interval $0 \leq t_A \leq 0.1$), then reaches a maximum value for t_A somewhere around 0.1, and, finally decreases monotonically in the interval $0.1 \leq t_A \leq 1$. The intuitive explanation for this behavior probably goes as follows. For $t_A = 0$, all customers belong to class B (with a short service time of 10 slots); as soon as t_A becomes positive, say $0 \leq t_A \leq 0.1$, most arriving customers are still of class B, but the sporadically arriving class-A customers (with a long service time of 100 slots), when in service, somehow block the regular processing of class-B customers. This effect causes the system content to increase. If, however, t_A increases further beyond the tipping point of $t_A = 0.1$ (while the total load ρ remains constant), the system starts receiving considerably less customers (of both types) for a given load, which explains the decreasing system content in the interval $0.1 \leq t_A \leq 1$.

The behavior of the mean waiting times, as shown in Fig. 9, is qualitatively a bit similar as for the mean system content. More specifically, it can be observed that the mean waiting times also increase for “low” values of t_A to reach a maximum value and then decrease for “higher” values of t_A . However, the maximum value of the waiting time is attained for t_A around 0.25, whereas the highest mean system content occurs for t_A in the vicinity of 0.1. Also, the rates at which the mean waiting times increase and decrease seem relatively slower than for the mean



Figure 8: Mean system content $E[s]$ versus t_A , for various values of ρ

system content. Intuitively, this can be attributed to the fact that the waiting time reflects the unfinished work in the system (at the arrival instant of a customer), while the system content indicates the number of customers in the system, whereby all customers contribute identically, irrespective of their service time, i.e., irrespective of the amount of work they represent. The fact that the mean waiting time (and, hence, the unfinished work in the system) for $t_A = 0$ is substantially smaller than for all other values of t_A can be explained by the higher burstiness of the arrival process of work units if class-B customers (bringing small amounts of work) are alternated with class-A customers (bringing large batches of work at the same time), which happens as soon as t_A gets positive.

7 Strong positive interclass correlation

The main results obtained so far for the general case are the formulas (21) and (22) for the pgf and the mean value of the system content at random slot boundaries. As discussed before, these formulas are expressed in terms of the basic parameters of our model, as well as the probabilities p_A and p_B that depend on the zero \hat{z} of the denominator of (21), which, in general, is to be determined numerically. In this section, we examine the behavior of the zero \hat{z} for high positive values of the interclass correlation γ . Let us first return to the first numerical example in section 6. Fig. 10 shows the zero \hat{z} for this example, as a function of the load ρ , for various “large”



Figure 9: Mean waiting time $E[w]$ versus t_A , for various values of ρ

values of γ . We observe that \hat{z} becomes closer and closer to the value 1 when $\gamma \rightarrow +1$, on condition that ρ remains smaller than its critical value 0.625. When $\rho \geq 0.625$, however, \hat{z} is not near to 1 at all.

Let us consider another example, for which the zero \hat{z} can be determined analytically. More specifically, let us assume Bernoulli arrivals (i.e., $E(z) = 1 - \lambda + \lambda z$), geometrically distributed service times (with mean $\mu_A > 1$) for class-A customers (as defined in equation (48)) and constant service times equal to 1 slot for class-B customers (i.e., $B(z) = z$ and $\mu_B = 1$). It is not difficult to see that, in this case, the denominator equation of (21) is equivalent to the following third-order polynomial equation in z :

$$(z - 1)\{(\rho_A - \lambda)[1 - \lambda + \lambda t_A(1 - \gamma)]z^2 - (1 - \lambda)[1 - \lambda\gamma + (\rho_A - \lambda)(1 - t_A(1 - \gamma))]z + \gamma(1 - \lambda)^2\} = 0 ,$$

where $\rho_A = \lambda\mu_A$, as defined in (49).

The above equation has one root $z = 1$ and two other roots z_1 and z_2 , solutions of the remaining quadratic equation, that can be expressed explicitly as follows:



Figure 10: The zero \hat{z} versus ρ , for high positive values of γ

$$z_1 = \frac{1 - \lambda}{2(\rho_A - \lambda)[1 - \lambda + \lambda t_A(1 - \gamma)]} [1 - \lambda\gamma + (\rho_A - \lambda)(1 - t_A(1 - \gamma)) + \sqrt{D}] ,$$

$$z_2 = \frac{1 - \lambda}{2(\rho_A - \lambda)[1 - \lambda + \lambda t_A(1 - \gamma)]} [1 - \lambda\gamma + (\rho_A - \lambda)(1 - t_A(1 - \gamma)) - \sqrt{D}] ,$$

where

$$D \triangleq [1 - \lambda\gamma + (\rho_A - \lambda)(1 - t_A(1 - \gamma))]^2 - 4\gamma(\rho_A - \lambda)[1 - \lambda + \lambda t_A(1 - \gamma)] . \quad (51)$$

Note that $z_1 \geq z_2$ for all feasible values of the system parameters.

Let us introduce some numerical values to study the behavior of the zeroes z_1 and z_2 . More specifically, let $t_A = 0.5$ (i.e., both customer classes are equiprobable) and $\mu_A = 5$. Then, $\rho_A = 5\lambda$ and $\rho = 3\lambda$, implying that global equilibrium is guaranteed as soon as $\lambda < 1/3$ or $\rho_A < 5/3$. Figs. 11 and 12 show the two zeroes z_1 and z_2 for this case, as functions of γ and ρ_A respectively.

Fig. 11 reveals that $z_1 > 1$ and $z_2 < 1$ for all feasible values of ρ_A , be it $\rho_A < 1$, $\rho_A = 1$ or $\rho_A > 1$, if γ is strictly less than 1 (which is implicitly assumed throughout this paper). This, of course, implies that the zero \hat{z} mentioned before is always given by $\hat{z} = z_2$. Fig. 12 further clarifies that $\hat{z} = z_2$ is very close to 1 for high positive values of γ , on condition that $\rho_A < 1$,



Figure 11: The zeroes z_1 and z_2 versus γ , for various values of ρ_A

but can differ substantially from 1 if $\rho_A > 1$. We note that the discriminant D , defined in (51), reduces to

$$D = (1 - 5\lambda)^2 = (1 - \rho_A)^2$$

if $\gamma \rightarrow +1$, so that

$$\sqrt{D} = 1 - \rho_A \quad , \quad \text{if } \rho_A < 1 \quad ,$$

$$\sqrt{D} = \rho_A - 1 \quad , \quad \text{if } \rho_A > 1 \quad ,$$

which explains the discontinuity in z_1 and z_2 at $\rho_A = 1$ in Fig. 12. These (analytical) results corroborate our conclusions from the previous (numerical) example. We now try to analyze the situation for arbitrary choices of $A(z)$, $B(z)$ and $E(z)$.

In general, the behavior of the zero \hat{z} for strong positive interclass correlation can be analyzed by considering the defining equation of \hat{z} , i.e., the requirement that the denominator of (21) should vanish, for $\gamma = 1$. With $\gamma = 1$, and hence $\alpha = \beta = 1$, this equation becomes :

$$[z - A(E(z))][z - B(E(z))] = 0 \quad .$$

In this particular case, the equation degenerates and is equivalent to two independent, simpler equations, i.e.,

$$z - A(E(z)) = 0$$



Figure 12: The zeroes z_1 and z_2 versus ρ_A , for high positive values of γ

and

$$z - B(E(z)) = 0 .$$

As in the previous examples, let us assume, without loss of generality, that class-A customers have a larger mean service time than class-B customers, i.e., that $\mu_A > \mu_B$. By means of Rouché's theorem [22, 8], it is not difficult to prove that when $\rho_A \triangleq \lambda\mu_A < 1$ and $\rho_B \triangleq \lambda\mu_B < 1$ (implying that the system is locally stable during A-periods and B-periods respectively), both equations have exactly one zero inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z -plane; moreover these zeroes are both equal to 1. This entails that, in this case, \hat{z} is also equal to 1. However, when $\rho_A > 1$ and $\rho_B < 1$, the first equation can be shown (again, by applying Rouché's theorem) to have two zeroes inside the closed unit disk, one of which is still equal to 1. The other zero lies strictly inside the unit disk and is the notorious zero \hat{z} that plays such an important role in the analysis of our system. The above makes plausible that, also in general, \hat{z} is close to 1 when $\gamma \rightarrow +1$ on condition that $\rho_A < 1$ and $\rho_B < 1$, but not necessarily if ρ_A increases beyond 1.

Let us assume now that both ρ_A and ρ_B are strictly lower than 1. We will show next that, in this case, the mean system content does not grow without bound when $\gamma \rightarrow +1$. In order to do so, we approximate γ by

$$\gamma \approx 1 - \epsilon , \tag{52}$$

where ϵ is a very small positive amount. As argued above, in these circumstances, the quantity

\hat{z} is also very close to 1, and, hence, can be approximated as

$$\hat{z} \approx 1 - r\epsilon + q\epsilon^2 , \quad (53)$$

where r and q are yet to be determined. In order to do so, we introduce the quantity ϵ in the defining equation of \hat{z} :

$$\hat{z}^2 - \hat{z}[\alpha A(E(\hat{z})) + \beta B(E(\hat{z}))] + \gamma A(E(\hat{z}))B(E(\hat{z})) = 0 .$$

Specifically, we make the substitutions (52) and (53) for γ and \hat{z} , we replace α and β by

$$\alpha \approx 1 - \epsilon(1 - t_A) ,$$

$$\beta \approx 1 - \epsilon t_A ,$$

respectively, and we approximate $A(E(\hat{z}))$ and $B(E(\hat{z}))$ as

$$A(E(\hat{z})) \approx 1 - \lambda\mu_A r\epsilon + [\lambda\mu_A q + A''(1)\frac{\lambda^2 r^2}{2} + \mu_A E''(1)\frac{r^2}{2}]\epsilon^2 ,$$

$$B(E(\hat{z})) \approx 1 - \lambda\mu_B r\epsilon + [\lambda\mu_B q + B''(1)\frac{\lambda^2 r^2}{2} + \mu_B E''(1)\frac{r^2}{2}]\epsilon^2 .$$

Identifying equal powers of ϵ on both sides of the resulting equation yields $0 = 0$ for the coefficients of ϵ^0 and ϵ^1 , but a useful equation for ϵ^2 , from which the quantity r appearing in (53) can be easily derived as

$$r = \frac{1 - \rho}{(1 - \rho_A)(1 - \rho_B)} . \quad (54)$$

Neglecting quadratic and higher powers of ϵ in (53), we can obtain the following first-order approximation for \hat{z} :

$$\hat{z} \approx 1 - \epsilon \frac{1 - \rho}{(1 - \rho_A)(1 - \rho_B)} ,$$

which is consistent with $\hat{z} < 1$ if $\rho_A < 1$ and $\rho_B < 1$ (and, hence, $\rho < 1$). Using the above result in the expressions (18) for p_A and p_B yields approximations for p_A and p_B which are independent of ϵ :

$$p_A \approx \frac{t_A(1 - \rho_A)}{1 - \rho} , \quad p_B \approx \frac{t_B(1 - \rho_B)}{1 - \rho} . \quad (55)$$

These can now be introduced in the last term of (47) to yield

$$\frac{\lambda(p_A - t_A)(\mu_A - \mu_B)}{1 - \gamma} \approx -\frac{t_A t_B \lambda^2 (\mu_A - \mu_B)^2}{\epsilon(1 - \rho)} .$$

Finally making use of this expression in (47) yields the following result (independent of ϵ and, hence, of γ) for the mean system content $E[s]$ in case the interclass correlation γ is close to +1 and both ρ_A and ρ_B are less than 1:

$$E[s] \approx \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} + \frac{t_A t_B \lambda^2 (\mu_A - \mu_B)^2}{1-\rho} . \quad (56)$$

This result explains the apparent contradiction between the analytic and numerical results that we observed earlier. Although equation (47) suggests that the mean system content $E[s]$ grows without bound for $\gamma \rightarrow +1$ *regardless of the other system parameters*, equation (56) reveals that this is not the case when both $\rho_A < 1$ and $\rho_B < 1$. It can be shown that in this case, the global stochastic equilibrium of the system performance measures (such as the mean system contents) can be obtained as the weighted average of these quantities in the two class-specific systems in equilibrium. These two class-specific systems correspond to considering the system evolution towards its (local) regime during the very long A -periods on the one hand, and the very long B -periods on the other hand. The (local) system regime during A -periods (or B -periods, respectively) can be analyzed with a single-class queueing model with service-time pgf $A(z)$ (or $B(z)$, respectively). These local regimes are characterized by the following pgf's of the system content *at customer departure times*:

$$\hat{P}_A(z) = \frac{(1-\rho_A)[E(z)-1]A(E(z))}{\lambda[z-A(E(z))]} ;$$

$$\hat{P}_B(z) = \frac{(1-\rho_B)[E(z)-1]B(E(z))}{\lambda[z-B(E(z))]} ,$$

so that the (global) pgf of the system content at arbitrary departure times (of either a class-A customer or a class-B customer) is given by

$$P(z) \approx t_A \hat{P}_A(z) + t_B \hat{P}_B(z) .$$

The corresponding pgf of the system content at random slot boundaries is [8, 7]

$$S(z) = P(z) \frac{\lambda(z-1)}{E(z)-1} .$$

Some straightforward calculations show that this is equivalent to

$$S(z) \approx (1-\rho)(z-1) \frac{z \left(\frac{t_A(1-\rho_A)}{1-\rho} A(E(z)) + \frac{t_B(1-\rho_B)}{1-\rho} B(E(z)) \right) - A(E(z))B(E(z))}{[z-A(E(z))][z-B(E(z))]} .$$

It is not difficult to see that exactly the same result is obtained when we replace γ by $+1$ and introduce the expression (55) for p_A and p_B in equation (21), obtained in subsection 3.3 for the pgf $S(z)$ in the general model studied in this paper. This proves that, indeed, the system exhibits some kind of quasi-stationary behavior when both $\rho_A < 1$ and $\rho_B < 1$ and $\gamma \rightarrow +1$. We label the global stochastic equilibrium of the system in this case with the term “strong” equilibrium or “strong” regime.

On the other hand, when the load $\rho < 1$, i.e., the system is (globally) stable, but $\rho_A > 1$ and $\rho_B < 1$, we refer to the global equilibrium of the system as a “compensated” equilibrium, because the overload periods created by the A -customers are *compensated* by the underload periods of the B -customers.

8 Conclusions and future work

In this paper, we have studied a dual-class, single-server queue in discrete time, operating under the global FCFS service discipline, assuming independent (aggregated) arrivals from slot to slot combined with a general first-order Markovian interclass correlation model. In our view, a first contribution of importance is the derivation of the main performance measures of the system in semi-analytical form, i.e., we have obtained explicit expressions for such quantities as the mean system content and the mean customer waiting time in terms of the basic parameters of the model and one parameter (\hat{z}) which is only implicitly known through a non-linear equation that it satisfies. Also, we have presented an analytic way to derive an expression for the steady-state pgf of the customer waiting time, that we consider to be a meaningful addition to existing work.

In addition to that, an extensive evaluation of these results has shown that the interclass correlation does not have any effect on the stability condition of the system, but it may have a very direct and great influence on the main performance measures of the system. More specifically, when the system is (globally) stable (i.e., when the global load ρ is strictly less than 1), we have observed that two different kinds of global equilibrium are possible, depending on the exact value of the load. For “low” values of the load (i.e., such that ρ_A and ρ_B are both lower than 1), the system exhibits a “strong” equilibrium, whereas for higher values of ρ (i.e., such that one of the quantities ρ_A and ρ_B is higher than 1 and the other lower than 1), the system reaches a “compensated” class of equilibrium. Especially in the latter case, the impact of strong positive interclass correlation may be devastating for the queueing performance. We consider these observations and conclusions to be another important contribution of the paper.

We believe that the phenomenon of class clustering in the context of multi-class queueing systems deserves more attention than it traditionally has received in the classical queueing literature. This is not only true for the system studied here, but also in other queueing situations whereby the service mechanism is sensitive to the order of service of customers of different classes. For instance, we have observed that class clustering may also have substantial effects on the performance of multi-class queues with multiple class-dedicated servers and a global FCFS service discipline [13, 15]. Also, in priority queues interclass correlation has been shown to have a possibly major impact on the delay-differentiating capabilities of the priority rule [31, 10, 11].

The model examined in this paper can be generalized in various directions. To start with, the assumption of independent aggregated arrivals from slot to slot may be relaxed to allow for correlated or bursty classes of arrival processes. We may also consider more complicated models for the interclass correlation in the arrival stream than the two-state Markovian model, e.g., the sizes of subsequent sequences of customers of each class could be described by general (rather than geometric) probability distributions, and so on. We plan to tackle several of these generalizations in future work. Our prior experience lets us believe that an analytic probability-generating functions approach may also work for these extended models, although we expect that generating numerical results will become increasingly involved, i.e., the number of zeroes inside the complex unit circle that must be calculated numerically will be larger than 1, and therefore the results will become harder to interpret and less intuitive to explain.

References

- [1] I. Adan and V. Kulkarni. Single-server queue with markov dependent inter-arrival and service times. *Queueing Systems*, 45(2):113–134, 2003.
- [2] I. Adan, A. Sleptchenko, and G. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research*, 26(4):559–585, 2009.
- [3] S. Asmussen. *Applied Probability and Queues*. Springer, 2002.
- [4] A. Banik, M. Chaudhry, and U. Gupta. On the finite buffer queue with renewal input and batch markovian service process: GI/BMSP/1/N. *Methodology and Computing in Applied Probability*, 10:559–575, 2008.
- [5] O. Boxma, J. Bruin, and B. Fralix. Sojourn times in polling systems with various service disciplines. *Performance Evaluation*, 66(11):621–639, 2009.
- [6] H. Bruneel. Queuing behavior of statistical multiplexers with correlated inputs. *IEEE Transactions on Communications*, 36(12):1339–1341, 1988.
- [7] H. Bruneel. Performance of discrete-time queuing-systems. *Computers & Operations Research*, 20(3):303–320, 1993.
- [8] H. Bruneel and B. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.
- [9] H. Bruneel, T. Maertens, B. Steyaert, D. Claeys, D. Fiems, and J. Walraevens. Analysis of a two-class FCFS queueing system with interclass correlation. In *Proceedings of the 19th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA '12)*, Grenoble, June 4-6 2012.
- [10] H. Bruneel, T. Maertens, and J. Walraevens. Interclass correlation in priority scheduling: an overlooked phenomenon. In *Proceedings of the Eighth International Conference on Queueing Theory and Network Applications (QTNA 2013)*, Taichung, Taiwan.
- [11] H. Bruneel, T. Maertens, and J. Walraevens. Class clustering destroys delay differentiation in priority queues. *European Journal of Operational Research*, 235(1):149–158, 2014.
- [12] H. Bruneel, W. Mélangé, B. Steyaert, D. Claeys, and J. Walraevens. Impact of blocking when customers of different classes are accommodated in one common queue. In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems (ICORES)*, Villamoura, Portugal, February 2012.
- [13] H. Bruneel, W. Mélangé, B. Steyaert, D. Claeys, and J. Walraevens. A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *European Journal of Operational Research*, 223:123–132, 2012.
- [14] N. Chrysos and M. Katevenis. Distributed WFQ scheduling converging to weighted max-min fairness. *Computer Networks*, 55:792–806, 2011.
- [15] D. Claeys, H. Bruneel, B. Steyaert, W. Mélangé, and J. Walraevens. Influence of data clustering on in-order multi-core processing systems. *Electronics Letters*, 49(1), 2012.

- [16] S. De Clercq, K. De Turck, B. Steyaert, and H. Bruneel. Frame-bound priority scheduling in discrete-time queueing systems. *Journal of Industrial and Management Optimization*, 7(3):767–788, 2011.
- [17] S. De Clercq, K. Laevens, B. Steyaert, and H. Bruneel. A multi-class discrete-time queueing system under the FCFS service discipline. *Annals of Operations Research*, 202(1):59–73, 2013.
- [18] S. De Vuyst, S. Wittevrongel, and H. Bruneel. Place reservation: Delay analysis of a novel scheduling mechanism. *Computers & Operations Research*, 35(8):2447–2462, 2008.
- [19] W. Feng and M. Umemura. Analysis of a finite buffer model with two servers and two nonpreemptive priority classes. *European Journal of Operational Research*, 192(1):151–172, 2009.
- [20] D. Fiems and H. Bruneel. A note on the discretization of Little’s result. *Operations Research Letters*, 30:17–18, 2002.
- [21] T. Frantti and M. Jutila. Embedded fuzzy expert system for adaptive weighted fair queueing. *Expert Systems with Applications*, 36(8):11390–11397, 2009.
- [22] M. González. *Classical complex analysis*. Marcel Dekker, New York, USA, 1992.
- [23] V. Gupta, M. Burroughs, and M. Harchol-Balter. Analysis of scheduling policies under correlated job sizes. *Performance Evaluation*, 67(11):996–1013, 2010.
- [24] A. Jeffrey. *Complex analysis and its applications*. CRC Press, London, 1992.
- [25] X. Jin and G. Min. Performance analysis of priority scheduling mechanisms under heterogeneous network traffic. *Journal of Computer and System Sciences*, 73(8):1207–1220, 2007.
- [26] X. Jin and G. Min. Analytical queue length distributions of GPS systems with long range dependent service capacity. *Simulation Modelling Practice and Theory*, 17(9):1500–1510, 2009.
- [27] M. Karsten. Approximation of generalized processor sharing with interleaved stratified timer wheels. *IEEE-ACM Transactions on Networking*, 18(3):708–721, 2010.
- [28] J. Kim, J. Kim, and B. Kim. Analysis of the M/G/1 queue with discriminatory random order service policy. *Performance Evaluation*, 68(3):256–270, 2011.
- [29] L. Kleinrock. *Queueing systems, part I*. Wiley, New York, USA, 1975.
- [30] P. Lieshout and M. Mandjes. Generalized processor sharing: Characterization of the admissible region and selection of optimal weights. *Computers & Operations Research*, 35(8):2497–2519, 2008.
- [31] T. Maertens, H. Bruneel, and J. Walraevens. Effect of class clustering on delay differentiation in priority scheduling. *Electronic Letters*, 48(10):568–569, 2012.
- [32] T. Maertens, J. Walraevens, and H. Bruneel. Performance comparison of several priority schemes with priority jumps. *Annals of Operations Research*, 180(3):1168–1185, 2008.
- [33] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys, and J. Walraevens. Impact of class clustering and global FCFS service discipline on the system occupancy of a two-class queueing model with two dedicated servers. In *7th International Conference on Queueing Theory and Network Applications, Proceedings*, 2012.

- [34] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys, and J. Walraevens. A continuous-time queueing model with class clustering and global FCFS service discipline. *Journal of Industrial Management and Optimization*, 10(1):193–206, 2014.
- [35] D. Min and Y. Yih. An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, 37(6):1091–1099, 2010.
- [36] J. Morrison. Two queues with vastly different arrival rates and processor-sharing factors. *Queueing Systems*, 64(1):49–67, 2010.
- [37] J. Morrison and S. Borst. Interacting queues in heavy traffic. *Queueing systems*, 65(2):135–156, 2010.
- [38] T. Ozawa. Analysis of queues with markovian service processes. *Stochastic Models*, 20(4):391–413, 2004.
- [39] B. Réveil, D. Claeys, T. Maertens, J. Walraevens, and H. Bruneel. Impact of class clustering in a multiclass fcfs queue with order-dependent service times. *Computers & Operations Research*, 51:90–98, 2014.
- [40] J. Shortle and M. Fischer. Approximation for a two-class weighted fair queueing discipline. *Performance Evaluation*, 67(10):946–958, 2010.
- [41] I. Verloop, U. Ayesta, and S. Borst. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems - Theory and Applications*, 20(4):473–509, 2010.
- [42] J. Walraevens, D. Fiems, and H. Bruneel. Time-dependent performance analysis of a discrete-time priority queue. *Performance Evaluation*, 65:641–652, 2008.
- [43] J. Walraevens, D. Fiems, S. Wittevrongel, and H. Bruneel. Calculation of output characteristics of a priority queue through a busy period analysis. *European Journal of Operational Research*, 198(3):891–898, 2009.
- [44] J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a GI-Geo-1 buffer with a preemptive resume priority scheduling discipline. *European Journal of Operational Research*, 157(1):130–151, 2004.
- [45] J. Walraevens, B. Steyaert, and H. Bruneel. A packet switch with a priority scheduling discipline: Performance analysis. *Telecommunication Systems*, 28(1):53–77, 2005.
- [46] J. Walraevens, J. van Leeuwen, and O. Boxma. Power series approximations for two-class generalized processor sharing systems. *Queueing systems*, 66(2):107–130, 2010.
- [47] L. Wang, G. Min, D. Kouvatsos, and X. Jin. Analytical modeling of an integrated priority and WFQ scheduling scheme in multi-service networks. *Computer Communications*, 33:S93–S101, 2010.
- [48] S. Zeltyn, Z. Feldman, and S. Wasserkrug. Waiting and sojourn times in a multi-server queue with mixed priorities. *Queueing Systems*, 61(4):305–328, 2009.
- [49] J. Zhao, B. Li, X. Cao, and I. Ahmad. A matrix-analytic solution for the DBMAP/PH/1 priority queue. *Queueing Systems*, 53(3):127–145, 2006.