ANALYSIS OF A FCFS QUEUE WITH TWO TYPES OF CUSTOMERS AND ORDER-DEPENDENT SERVICE TIMES

Bert Réveil, Dieter Claeys, Tom Maertens, Joris Walraevens, Herwig Bruneel
SMACS Research Group, TELIN Department
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
Email: {breveil, dclaeys, tmaerten, jw, hb}@telin.ugent.be

KEYWORDS

Queueing; Order-dependent service times; Class clustering

ABSTRACT

In this paper, we study a discrete-time first-come-first-served queueing system with a single server and two types (classes) of customers, where the (average) service time of a customer is longer if its type differs from the type of the preceding customer. As opposed to traditional literature, the different types of customers do not occur randomly and independently in the arrival stream: we include a Markovian type of correlation in the types of consecutive customers instead. We deduce the probability generating function of the system content, from which we extract various performance measures, such as the mean values of the system content and the customer delay. We demonstrate that the interclass correlation in the arrival stream has a tremendous impact on the system performance, which highlights the necessity to include it in the performance assessment of the system.

I. INTRODUCTION

In this paper, we study a discrete-time queueing system with two types (classes) of customers, a common queue, and one server. Customers are served in order of arrival, i.e., the queueing discipline is first-come-first-served (FCFS), irrespective of the class the consecutive customers belong to (referred to as "global FCFS" in this paper). However, the service time of a customer depends on the identity or non-identity of its class and the class of the preceding customer. Specifically, we assume that the (average) service time of any customer is longer if its type differs from the type of the previously served customer, a feature that arises regularly in practice. One major reason for this phenomenon may be the necessity to reconfigure or adapt the service facility for other tasks than the current one. Our model also applies to many other situations. Customers of distinct types could correspond, for instance, to vehicles that are heading to other destinations at road intersections, jobs with different execution times (because they require other resources), people requiring distinct kinds of services at a call center, (semi-finished) products that need different machines to be processed, printing jobs in a specialized printing house that delivers print work in several different formats, different types of goods being delivered in warehouses and being stocked in different sections, etc. The common feature of all these applications is that the service time of the next customer is, on average, longer if the preceding and the next customer (vehicle, job, person, product, printing job, goods delivery) require a different kind of service, i.e., do not belong to the same (service) class.

In traditional research on multi-class queues (see e.g. [1], [3], [8], [9], [13], [14], [18], [19]) it is standard to assume that the different types of customers occur randomly and independently in the arrival stream of customers into the system, which is often in contrast to the actual situation. In reality, there is usually some degree of *interclass correlation* or *class clustering*. In some cases, for instance, customers of the same type have a tendency to arrive "back to back". As an example, consider a network router transmitting data from and towards various communicating processes running in the network. Within certain time frames of its service, it is likely that the router will transfer consecutive data packets that all originate from the same process.

In a number of recent papers [6], [7], [16], [15], [5], we have revealed that class clustering can have a major impact on the performance of several other queueing systems with two types of customers, such as queues with multiple class-dedicated servers and global FCFS service [6], [7], [16], priority queues [15], and single-server queues with global FCFS and class-dependent service times (regardless of the type of the previous customer) [5]. Therefore, we presume that this will also hold in the queueing system under investigation in the present paper, i.e., a system where the mean service time of a customer is longer when its type differs from the type of the previously served customer.

The paper is structured as follows. First, we describe the system in section II. Then, in section III,

Proceedings 28th European Conference on Modelling and Simulation ©ECMS Flaminio Squazzoni, Fabio Baronio, Claudia Archetti, Marco Castellani (Editors)
ISBN: 978-0-9564944-8-1 / ISBN: 978-0-9564944-9-8 (CD)

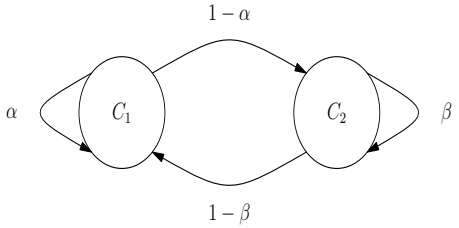

Fig. 1. Two-state Markov chain of the customer types.

we analyze the system behavior and we establish the probability generating function (PGF) of the number of customers in the system, hereafter referred to as the *system content*, both at customer departure times and at random slot boundaries. Next, the influence of class clustering is investigated in section IV, and finally, some conclusions are drawn and indications for further research given in section V.

II. SYSTEM DESCRIPTION

We study a discrete-time queueing system with an infinite waiting room, one server, and two types (classes) of customers, named C_1 and C_2. The time axis is divided into fixed-length time intervals, referred to as *slots* in the sequel. New customers can arrive in the system at any given (continuous) point on the time axis, but customer service times can only start and end at slot boundaries. Customers are served according to a *global FCFS* service discipline, meaning that they are served in order of arrival, regardless of the class they belong to.

The arrival process of new customers is characterized in two steps. First, the total (aggregated) number of customer arrivals in consecutive slots is represented by a sequence of independent and identically distributed (IID) nonnegative discrete random variables with common probability mass function (PMF) $e(n)$ and probability generating function (PGF) $E(z)$:

$$\begin{aligned} e(n) &\triangleq \text{Prob}[n \text{ arrivals in one slot}] \ , \ n \geq 0 \\ E(z) &\triangleq \sum_{n=0}^{\infty} e(n) z^n . \end{aligned}$$

The *(total) mean arrival rate*, i.e., the (total) mean number of arrivals per slot, is given by

$$\lambda \triangleq E'(1). \tag{1}$$

Secondly, the occurrence of type C_1 and type C_2 customers within the total arrival stream is governed by a customer-type correlation model. This implies that we account for the possibility of *interclass correlation*, or *class clustering* in the arrival process. Customers of any given type may (or may not) have a tendency to "arrive back-to-back". Consequently, the types of consecutive customers may be non-independent. In this study, we consider a first-order Markovian type of correlation between the types of consecutive customers (see Fig. 1).

If t_k denotes the type of customer k, the transition probabilities of the Markov chain that determines the types of consecutive customers are defined as

$$\begin{aligned} \alpha &\triangleq \text{Prob}[t_{k+1} = C_1 | t_k = C_1] \ , \\ \beta &\triangleq \text{Prob}[t_{k+1} = C_2 | t_k = C_2] \ . \end{aligned} \tag{2}$$

The steady-state probabilities t_{C_1} and t_{C_2} of finding the Markov chain in state C_1 respectively C_2 are given by [10], [12]

$$\begin{aligned} t_{C_1} &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = C_1] = \frac{1-\beta}{2-\alpha-\beta} \ , \\ t_{C_2} &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = C_2] = \frac{1-\alpha}{2-\alpha-\beta} \ . \end{aligned} \tag{3}$$

They can be interpreted as the fractions of type C_1 and type C_2 customers in the arrival stream. Defining T_k as a numerical variable obeying

$$T_k = 1 \iff t_k = C_1 \ , \text{ and } T_k = 0 \iff t_k = C_2 \ ,$$

the steady-state correlation coefficient γ $(-1 \leq \gamma \leq 1)$ of the Markov chain, called the *interclass correlation* in the sequel, is defined as

$$\begin{aligned} \gamma &\triangleq \lim_{k\to\infty} \frac{\text{E}\left[T_k T_{k+1}\right] - \text{E}\left[T_k\right] \text{E}\left[T_{k+1}\right]}{\sqrt{\text{var}[T_k] \ \text{var}[T_{k+1}]}} \\ &= \alpha + \beta - 1 \ . \end{aligned} \tag{4}$$

It represents the amount of correlation between the types of two consecutive customers in the arrival stream (in the steady-state). Positive values of γ correspond to situations in which at least one customer type has a tendency to cluster. Negative values of γ typically imply (strongly) alternating customer type arrivals. If $\gamma = 0$, and consequently $\alpha = 1 - \beta$, the types of consecutive customers are independent, corresponding to the situation that is traditionally (implicitly) assumed in literature.

The *service time* of a customer indicates the number of slots needed to fully serve that customer. We assume that the service time of a customer depends on its own type and on the type of the previous customer. If both types are the same, the service time equals one slot, whereas in the other case, the service time is strictly positive and characterized by the PMF $b(n)$ $(n \geq 1)$, PGF

$$B(z) \triangleq \sum_{n=1}^{\infty} b(n) z^n \ ,$$

and mean value

$$\mu_B^{-1} \triangleq B'(1) > 1 \ . \tag{5}$$

III. SYSTEM ANALYSIS

In this section, we first present an analysis of the total number of customers in the system at customer departure times. The PGF is established under steady-state conditions and a method is described to determine the two remaining unknowns in the expression. Finally, we deduce the PGF and the average value of the system content at random slot boundaries.

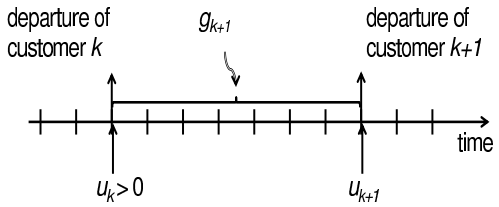

Fig. 2. Relationship between u_k and u_{k+1} when $u_k > 0$.

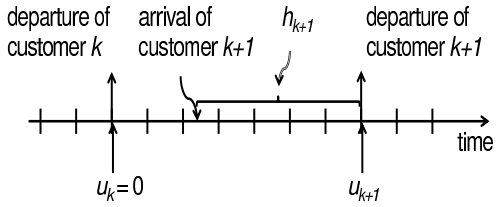

Fig. 3. Relationship between u_k and u_{k+1} when $u_k = 0$.

System equations at customer departure times

In this subsection, we establish system equations that capture the behavior of the system content at customer departure times. To this end, let u_k represent the total number of customers in the system immediately after the service completion of the k-th customer. Due to the assumptions presented in section II, the sequence of couples (t_k,u_k) constitutes a first-order Markov chain describing the evolution of the system from slot to slot. As described above, the state transitions for the sequence $\{t_k\}$ are governed by Equation (2). For the quantities $\{u_k\}$, we obtain two recursive equations that cover the complete set of situations depicted in figures 2 and 3:

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

In these equations, g_{k+1} stands for the (total) number of arrivals during the service time of customer $k+1$. The quantity h_{k+1} can be written as

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

with f_{k+1} the number of customer arrivals in the arrival slot of customer $k+1$, but *after* customer $k+1$ (in case customer $k+1$ enters an empty system).

Its PMF $f(n)$ and PGF $F(z)$ can be found as

$$\begin{aligned} f(n) &\triangleq \text{Prob}[n \text{ additional arrivals}|\text{at least 1 arrival}] \\ &= \frac{e(n+1)}{1-E(0)} \text{ , } n \geq 0 \text{ ,} \\ F(z) &\triangleq E[z^{f_{k+1}}] = \sum_{n=0}^{\infty} f(n) z^n \\ &= \frac{E(z)-E(0)}{z[1-E(0)]} \text{ ,} \end{aligned} \tag{7}$$

irrespective of whether customer $k+1$ is of the same or different type as customer k. For the PGFs of the quantities g_{k+1} and h_{k+1}, the equality of the customer types of two consecutive customers does make a difference. Taking into account that we are considering an IID aggregated arrival process, implying that f_{k+1} and g_{k+1} are mutually independent, we find that

$$\begin{aligned} \mathrm{E}\left[z^{g_{k+1}}|t_{k+1} = t_k\right] &= E(z) \text{ ,} \\ \mathrm{E}\left[z^{h_{k+1}}|t_{k+1} = t_k\right] &= F(z)E(z) \text{ ,} \\ \mathrm{E}\left[z^{g_{k+1}}|t_{k+1} \neq t_k\right] &= B(E(z)) \text{ ,} \\ \mathrm{E}\left[z^{h_{k+1}}|t_{k+1} \neq t_k\right] &= F(z)B(E(z)) \text{ .} \end{aligned}$$

System content at customer departure times

One of our intentions is to provide expressions for the performance measures of the queueing system under steady-state conditions. It is well-known [4], [17] that for any work-conserving queueing system, stability is guaranteed when the average amount of work entering the system per slot (often referred to as the *work load* ρ) is strictly less than the amount of work that can be delivered by the server per slot. In our model, considering a single server without interruptions, the stability condition thus boils down to

$$\rho \triangleq \lambda \mathrm{E}[s] < 1 \text{ ,}$$

with s the steady-state service time of an arbitrary customer. Using the law of the total expectation, $\mathrm{E}[s]$ can be expanded, yielding

$$\mathrm{E}[s] = t_A + t_B \mu_B^{-1} \text{ ,} \tag{8}$$

where $t_A = \alpha t_{C_1} + \beta t_{C_2}$ and $t_B = (1-\alpha)t_{C_1} + (1-\beta)t_{C_2}$ denote the steady-state probabilities that two consecutive customers belong to the same or the opposite class respectively. If we rework Equation (8), substituting α and β in terms of γ, t_{C_1} and t_{C_2}, another, more interesting expression for ρ can be found that links the work load directly to the amount of interclass correlation in the arrival process:

$$\rho = \lambda[1 + 2(1-\gamma)(\mu_B^{-1} - 1)t_{C_1}t_{C_2}] \text{ .} \tag{9}$$

As could have been anticipated, we find that in case of ultimate positive customer type correlation ($\gamma = 1$, i.e., both $\alpha = 1$ and $\beta = 1$), the work load reduces to λ. More generally, this also holds for single-class systems where either α or β is equal to 1, because in that case either t_{C_1} or t_{C_2} equals 0. If γ equals -1, i.e., if the customer type changes with every customer arrival, t_{C_1} and t_{C_2} are both equal to 0.5, and the work load increases to $\lambda\mu_B^{-1}$, also as expected.

Assuming that the stability condition is met, we define joint steady-state probabilities for the Markov chain $\{(t_k, u_k)\}$ as

$$\begin{aligned} p_{C_1}(i) &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = C_1, u_k = i] \text{ ,} \\ p_{C_2}(i) &\triangleq \lim_{k\to\infty} \text{Prob}[t_k = C_2, u_k = i] \text{ ,} \end{aligned}$$

for all $i \geq 0$. The corresponding partial PGFs are given by

$$P_{C_1}(z) \triangleq \sum_{i=0}^{\infty} p_{C_1}(i) z^i \text{ , } P_{C_2}(z) \triangleq \sum_{i=0}^{\infty} p_{C_2}(i) z^i \text{ ,}$$

while the steady-state PGF $P(z)$ of the total system content at customer departure times is given by

$$P(z) = P_{C_1}(z) + P_{C_2}(z) \ . \tag{10}$$

Relying on the balance equations of the Markov chain, it is now possible to establish two linear independent equations for the partial PGFs $P_{C_1}(z)$ and $P_{C_2}(z)$. For customers of class C_1, we get

$$\begin{aligned} p_{C_1}(j) &= \lim_{k\to\infty} \text{Prob}[t_{k+1} = C_1,\, u_{k+1} = j] \\ &= \sum_{i=0}^{\infty} \lim_{k\to\infty} \text{Prob}[t_k = C_1,\, u_k = i] \\ &\quad \text{Prob}[t_{k+1} = C_1,\, u_{k+1} = j | t_k = C_1,\, u_k = i] \\ &\quad + \sum_{i=0}^{\infty} \lim_{k\to\infty} \text{Prob}[t_k = C_2,\, u_k = i] \\ &\quad \text{Prob}[t_{k+1} = C_1,\, u_{k+1} = j | t_k = C_2,\, u_k = i] \\ &= \alpha \sum_{i=0}^{\infty} p_{C_1}(i) \lim_{k\to\infty} \qquad (11) \\ &\quad \text{Prob}[u_{k+1} = j | u_k = i,\, t_k = C_1,\, t_{k+1} = C_1] \\ &\quad + (1-\beta) \sum_{i=0}^{\infty} p_{C_2}(i) \lim_{k\to\infty} \\ &\quad \text{Prob}[u_{k+1} = j | u_k = i,\, t_k = C_2,\, t_{k+1} = C_1] \ . \end{aligned}$$

Taking the z-transform of (11) yields:

$$\begin{aligned} P_{C_1}(z) &\triangleq \sum_{j=0}^{\infty} p_{C_1}(j) z^j \\ &= \alpha \sum_{i=0}^{\infty} p_{C_1}(i) \lim_{k\to\infty} \\ &\quad \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = C_1,\, t_{k+1} = C_1\right] \quad (12) \\ &\quad + (1-\beta) \sum_{i=0}^{\infty} p_{C_2}(i) \lim_{k\to\infty} \\ &\quad \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = C_2,\, t_{k+1} = C_1\right] . \end{aligned}$$

The expectations in the above equations can be elaborated using the system equations in (6):

$$\begin{aligned} &\lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = C_1,\, t_{k+1} = C_1\right] \\ &= \lim_{k\to\infty} \text{E}\left[z^{i-1+g_{k+1}} | t_k = C_1,\, t_{k+1} = C_1\right] \\ &\qquad = z^{i-1} E(z) \ , \text{ for all } i \geq 1 \ , \qquad (13) \\ &\lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = 0,\, t_k = C_1,\, t_{k+1} = C_1\right] \\ &= \lim_{k\to\infty} \text{E}\left[z^{h_{k+1}} | t_k = C_1,\, t_{k+1} = C_1\right] \\ &\qquad = F(z)E(z) \ , \text{ for } i = 0 \ , \qquad (14) \\ &\lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = C_2,\, t_{k+1} = C_1\right] \\ &= \lim_{k\to\infty} \text{E}\left[z^{i-1+g_{k+1}} | t_k = C_2,\, t_{k+1} = C_1\right] \\ &\qquad = z^{i-1} B(E(z)) \ , \text{ for all } i \geq 1 \ , \qquad (15) \\ &\lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = 0,\, t_k = C_2,\, t_{k+1} = C_1\right] \\ &= \lim_{k\to\infty} \text{E}\left[z^{h_{k+1}} | t_k = C_2,\, t_{k+1} = C_1\right] \\ &\qquad = F(z)B(E(z)) \ , \text{ for } i = 0 \ . \qquad (16) \end{aligned}$$

Substitution of (13), (14), (15) and (16) in expression (12) finally leads to a first linear equation between $P_{C_1}(z)$ and $P_{C_2}(z)$:

$$\begin{aligned} (z - \alpha E(z)) P_{C_1}(z) &= (1-\beta) B(E(z)) P_{C_2}(z) \\ &\quad + \alpha E(z)(zF(z) - 1) P_{C_1}(0) \qquad (17) \\ &\quad + (1-\beta) B(E(z))(zF(z) - 1) P_{C_2}(0) \ . \end{aligned}$$

Similarly, a second linear equation can be found starting from the balance equations for type C_2 customers:

$$\begin{aligned} (z - \beta E(z)) P_{C_2}(z) &= (1-\alpha) B(E(z)) P_{C_1}(z) \\ &\quad + \beta E(z)(zF(z) - 1) P_{C_2}(0) \qquad (18) \\ &\quad + (1-\alpha) B(E(z))(zF(z) - 1) P_{C_1}(0) \ . \end{aligned}$$

Equations (17) and (18) can be solved for the unknown partial PGFs $P_{C_1}(z)$ and $P_{C_2}(z)$. Using the results and Equation (7) to expand Equation (10), we obtain a first expression for the PGF $P(z)$:

$$\begin{aligned} P(z) = \frac{P(0)(E(z)-1)}{1 - E(0)} \times \qquad (19) \\ \frac{z(p_A E(z) + p_B B(E(z))) - \alpha\beta E(z)^2 + (1-\alpha)(1-\beta) B(E(z))^2}{z^2 - z(\alpha+\beta)E(z) + \alpha\beta E(z)^2 - (1-\alpha)(1-\beta) B(E(z))^2} \ , \end{aligned}$$

with

$$\begin{aligned} p_A &\triangleq \frac{\alpha P_{C_1}(0) + \beta P_{C_2}(0)}{P(0)} \ , \\ p_B &\triangleq \frac{(1-\alpha) P_{C_1}(0) + (1-\beta) P_{C_2}(0)}{P(0)} \ . \qquad (20) \end{aligned}$$

Given that $P(0) = P_{C_1}(0) + P_{C_2}(0)$, these quantities are equal to

$$\begin{aligned} p_A &= \lim_{k\to\infty} \text{Prob}[t_{k+1} = t_k | u_k = 0] \ , \\ p_B &= \lim_{k\to\infty} \text{Prob}[t_{k+1} \neq t_k | u_k = 0] \ , \qquad (21) \end{aligned}$$

the conditional probabilities that a new customer entering an empty system (in steady state) is of the same or the opposite type respectively as the last customer that was served by the system.

Expression (19) still contains three unknowns that need to be determined: $P(0)$, p_A and p_B. The probability $P(0)$ can be found by imposing the normalization condition on the PGF $P(z)$, i.e. $P(1) = 1$. Using de l'Hôpital's rule to solve the equation, we obtain that

$$P(0) = \frac{(1 - E(0))(1-\rho)}{\lambda} \ . \tag{22}$$

In order to derive expressions for p_A and p_B, two linear equations in p_A and p_B are established. The first one simply states that

$$p_A + p_B = 1 \ . \tag{23}$$

The second equation follows from the fact that a PGF, like $P(z)$, is bounded inside the closed unit disk of the complex z-plane $\{z \in \mathbb{C} : |z| \leq 1\}$. As it can be proved via Rouché's theorem [2] that the denominator of $P(z)$ has exactly two zeroes inside the closed unit disk, the aforementioned property of PGFs implies that those zeroes, one of which is equal to 1, must also be zeroes of $P(z)$'s numerator. Otherwise, the function value would

tend to infinity inside the closed unit z-disk. For $z = 1$, given its factor $(E(z) - 1)$, the numerator clearly vanishes. For the second zero however, called $\hat{z}$ from here on, the other factor in the numerator should equal 0, which yields a linear equation for p_A and p_B. Solving this equation, in combination with Equation (23), we find that p_A and p_B can be determined as

$$\begin{aligned} p_A &= \frac{(\alpha+\beta)E(\hat{z}) - B(E(\hat{z})) - \hat{z}}{E(\hat{z}) - B(E(\hat{z}))} , \\ p_B &= \frac{(1-\alpha-\beta)E(\hat{z}) + \hat{z}}{E(\hat{z}) - B(E(\hat{z}))} . \end{aligned} \tag{24}$$

Once the zero $\hat{z}$ is computed numerically, e.g. via standard root-finding functions in Maple or Matlab, p_A and p_B are fixed, and as such, so is $P(z)$.

System content at random slot boundaries

From earlier research [4], it follows that for all discrete-time single-server queueing systems with (general) independent customer arrivals from slot to slot (with PGF $E(z)$), a fairly simple relationship holds between the PGF $P(z)$ of the system content at customer departure times and the PGF $U(z)$ of the system content at random slot boundaries, regardless of the exact characteristics of the service process and the intra-slot details of the arrival process (e.g., single or batch arrivals, the exact time customers arrive within the slot, etc.). This relationship is

$$P(z) = \frac{E(z)-1}{\lambda(z-1)} U(z) . \tag{25}$$

As the examined model belongs to the class of systems described above, relationship (25) in combination with Equations (19) and (22) leads to the following expression for the PGF of the system content at random slot boundaries:

$$U(z) = (1-\rho)(z-1)\times \frac{z(p_A E(z) + p_B B(E(z))) - \alpha\beta E(z)^2 + (1-\alpha)(1-\beta)B(E(z))^2}{z^2 - z(\alpha+\beta)E(z) + \alpha\beta E(z)^2 - (1-\alpha)(1-\beta)B(E(z))^2} . \tag{26}$$

From this expression, various interesting performance measures can be derived, one of which is the mean system content $\mathrm{E}[u]$ at random slot boundaries. The latter can be determined as $\mathrm{E}[u] = U'(1)$, where u represents the system content at the beginning of a random slot in steady state. After long and tedious calculations, we find that

$$\begin{aligned} \mathrm{E}[u] &= \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} \\ &\quad + 2\lambda(1-\mu_B^{-1})t_{C_1}t_{C_2} + \frac{p_B\lambda(\mu_B^{-1}-1)}{1-\gamma} \\ &\quad + \frac{(1-\gamma)\lambda^2(\mu_B^{-1}-1)^2 t_{C_1}t_{C_2}(1-4t_{C_1}t_{C_2})}{1-\rho} , \end{aligned} \tag{27}$$

with $C'(1)$ and $C''(1)$ the first two derivatives for $z = 1$ of the PGF $C(z)$ of the service time of an arbitrary customer:

$$C(z) = t_A z + (1-t_A)B(z) , \tag{28}$$

with

$$t_A = \lim_{k\to\infty} \mathrm{Prob}[\mathrm{t_k} = \mathrm{t_{k-1}}] .$$

In Equation (27), the first term ρ accounts for the average server content, or the mean number of customers in service. The last four terms cover the mean queue occupancy, meaning the average number of customers that are waiting to be served.

Higher-order moments of the system content at random slot boundaries can be obtained by computing higher-order derivatives of the PGF $U(z)$. By means of Little's law (for discrete-time queues) [11], one can determine the average *delay* (system time) of an arbitrary customer as $\mathrm{E}[d] = \mathrm{E}[u]/\lambda$ (d stands for the delay of a random customer in the system in steady state). The mean of the *waiting time* w is obtained as $\mathrm{E}[w] = \mathrm{E}[d] - \mathrm{E}[s]$, where $\mathrm{E}[s]$ was defined in (8). In our case, it is given by

$$\begin{aligned} \mathrm{E}[w] &= \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2\lambda(1-\rho)} \\ &\quad + 2(1-\mu_B^{-1})t_{C_1}t_{C_2} + \frac{p_B(\mu_B^{-1}-1)}{1-\gamma} \\ &\quad + \frac{(1-\gamma)\lambda(\mu_B^{-1}-1)^2 t_{C_1}t_{C_2}(1-4t_{C_1}t_{C_2})}{1-\rho} . \end{aligned}$$

IV. DISCUSSION OF RESULTS AND NUMERICAL EXAMPLES

In this section, we discuss the obtained results, both from a qualitative perspective as by means of some numerical examples. The first interesting result was already given by Equation (9). The equation expresses the direct dependency of the work load ρ on the inter-class correlation factor γ ($\triangleq \lambda\mathrm{E}[s]$). Consequently, the stability condition,

$$\lambda < \frac{1}{\mathrm{E}[s]} = \frac{1}{1 + 2(1-\gamma)(\mu_B^{-1}-1)t_{C_1}t_{C_2}} , \tag{29}$$

reveals that the supremum of the achievable throughput of the presented system, denoted as λ_{sup}, and expressed in customers per slot, depends on γ, μ_B^{-1} and the fractions of C_1 and C_2 customers.

Equation (29) reveals that if μ_B^{-1} increases, λ_{sup} decreases, because the mean service time for customers following customers of the opposite type is increased.

For fixed μ_B^{-1}, we find that λ_{sup} is lowest when $t_{C_1}t_{C_2}$ reaches it maximal value, i.e., for $t_{C_1} = t_{C_2} = \frac{1}{2}$. If one type of customers enters the system more often than the other ($t_{C_1} > 0.5 > t_{C_2}$ or $t_{C_2} > 0.5 > t_{C_1}$), consecutive customers will be of the same type more often, implying that the average service time of an arbitrary customer decreases, or that the throughput of the system increases.

When t_{C_1}, t_{C_2} and μ_B^{-1} are fixed, the throughput decreases when the types of consecutive customers begin to alter more regularly, i.e., when γ becomes smaller. The worst case occurs when $\gamma = -1$ meaning that the types of subsequent customers always differ. The best scenario occurs when only one type of customers enters the system ($\gamma = 1$).

A second interesting result was given by Equation (27). This expression clearly indicates the influence of the different system parameters on the mean system content at random slot boundaries. The first two terms of Equation (27) correspond to the classical terms that constitute the expression for the average system content at random slot boundaries of a system with no interclass correlation and a service-time PGF $C(z)$. The other three terms in the expression can be fully attributed to the presence of class clustering in the arrival process.

It is not surprising to see that the mean system content depends on the first two moments of the aggregated arrival process (represented by the quantities λ, $E''(1)$ and $\rho = \lambda E[s]$) and on the first two moments of the service times (represent by the quantities $C'(1)$, $C''(1)$, μ_B^{-1} and $\rho = \lambda C'(1)$). Furthermore, we anticipated to find that the mean system content goes to infinity as soon as the work load ρ approaches its limiting value 1.

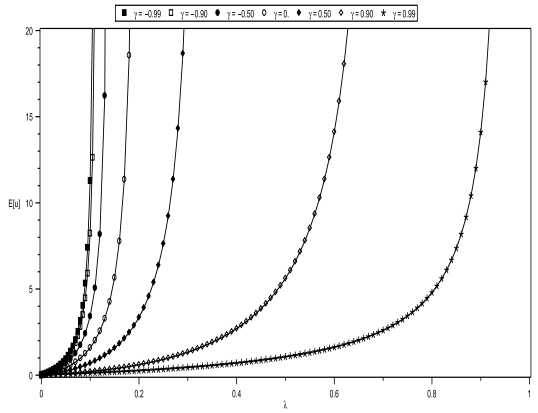

Fig. 4. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $B(z)$ given by (30), $\mu_B^{-1} = 9$, $t_{C_1} = t_{C_2} = 0.5$ and several interclass correlation factors.

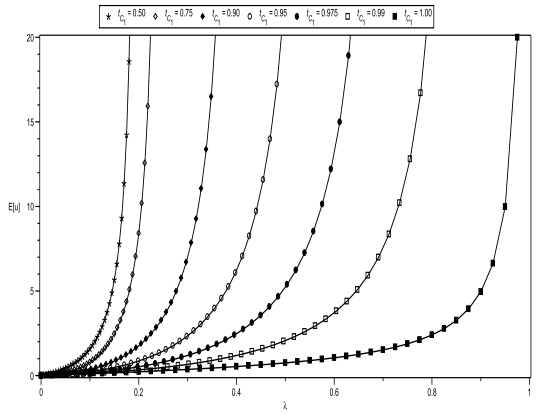

Fig. 5. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $B(z)$ given by (30), $\mu_B^{-1} = 9$, $\gamma = 0$ and various fractions of customer types in the arrival stream.

In Figures 4-6, we present numerical results for two-class queueing systems dealing with an aggregated Poisson arrival process (i.e., $E(z) = e^{\lambda(z-1)}$) and the following PGF of the service time of customers following a customer of the opposite type:

$$B(z) = \frac{z}{\mu_B^{-1} + (1 - \mu_B^{-1})z} \quad . \tag{30}$$

Figure 4 shows the mean system content versus λ for different values of γ, in a system where $\mu_B^{-1} = 9$ and both types of customers occur with the same a priori frequency (i.e., $t_{C_1} = t_{C_2} = 0.5$). The average number of customers the system can deal with depends heavily on the amount of interclass correlation: the more positive, the more customers can be served per slot. This implies that the system occupancy raises rapidly for systems with a negative interclass correlation.

In Figure 5, we examine the impact of the fractions of type C_1 and type C_2 customers in the arrival stream on the average system content. The figure depicts the mean system content versus λ, in a system where $\mu_B^{-1} = 9$, and with a fixed interclass correlation of 0.

The figure mainly shows that having two types of customers instead of one, strongly affects the mean system content. If only one type of customer occurs, the average system content is much lower, because every arriving customer only requires one time slot to be served. As soon as two different types of customers enter the system, the average system content increases considerably. As reasoned before, based on Equation (29), the exact fraction of type C_1 and type C_2 customers influences the achievable throughput of the system.

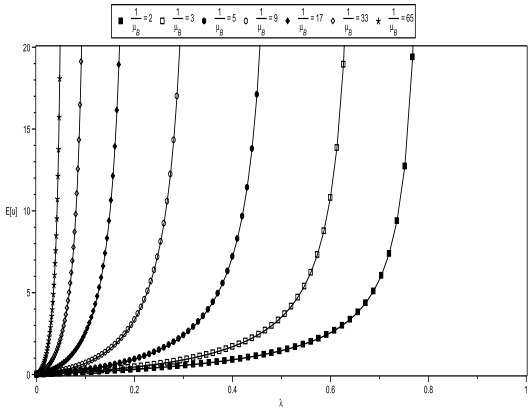

Fig. 6. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $\gamma = 0.5$ and $t_{C_1} = t_{C_2} = 0.5$; $B(z)$ given by (30).

In a third plot (Figure 6), we present the mean system content of a system that is facing a positive interclass correlation of 0.5 and an equal amount of type C_1 and type C_2 customers. The mean service time μ_B^{-1} is varied. The average system content increases when μ_B^{-1} increases. If the interclass correlation factor is fixed, a longer service time for customers that are not of the same type as the previous customer implies more arriving customers during that service time, and consequently, more customers waiting in the system.

V. CONCLUSIONS AND FURTHER RESEARCH

We have investigated a discrete-time queueing system with two types (classes) of customers, one server,

a common queue, global FCFS queueing discipline, and service times that are, on average, longer when the types of consecutive customers differ. The aim of the paper was to study the impact of class clustering on the system performance, a feature that is traditionally overlooked in the literature. As we have already revealed that class clustering has a major impact on other systems, we presumed that this statement also holds for a system with order-dependent service times. We have therefore studied the performance, both in terms of stability and system content, in such a system that is subject to class clustering. We have found that class clustering has an undeniable impact on the system performance. This conclusion highlights the necessity to incorporate class clustering when studying a queue with order-dependent service times. Our results can be used for that purpose.

This paper is a first (conceptual) step in a more general study of queueing systems with order-dependent service times. A first extension could be a model in which the service-time distribution of a customer is not only dependent on the identity or non-identity of its type and the previous customer's type, but also on the actual type itself. In the present paper, the service times are deterministic and equal to one slot when the customer at hand has the same type as the preceding customer, while the service time is, on average, longer than one slot in the opposite case; this could be generalized such that in both cases the service-time distributions are different and completely arbitrary. Future research may also consider non-Markovian types of interclass correlation in the arrival stream, e.g., arrival processes where the numbers of consecutive customers of the same type have more general probability distributions than the geometric distributions considered in this paper. Another interesting extension is to incorporate slot-to-slot correlation in the (aggregated) arrival process, both with respect to the total numbers of arrivals per slot and the types of customers arriving in different slots. A generalization to more customer types seems complicated, especially when the state transitions of the modulating Markov chain of the arrival process are completely arbitrary. A matrix-analytic approach may be more feasible here, but the derivation of explicit closed-form results could be harder. Finally, it may be interesting to investigate related continuous-time queueing models. Some of these issues will be dealt with in the future.

References

[1] I.J.B.F. Adan, A. Sleptchenko, and G.J. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research*, 26(4):559–585, 2009.

[2] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[3] M.A.A. Boon, I.J.B.F. Adan, and O.J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67:468–484, 2010.

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

[5] 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.

[6] H. Bruneel, W. Mélange, 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.

[7] H. Bruneel, W. Mélange, 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.

[8] H. Chen and H. Zhang. Stability of multiclass queueing networks under priority service disciplines. *Operations Research*, 48(1):26–37, 2000.

[9] E.B. Cil, F. Karaesmen, and E.L. Ormeci. Dynamic pricing and scheduling in a multi-class single-server queueing system. *Queueing Systems*, 67(4):305–331, 2011.

[10] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, Third Edition.* Wiley, New York, 1968.

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

[12] R.G. Gallager. *Discrete stochastic processes.* Kluwer Academic, Boston/Dordrecht/London, 1996.

[13] D. Gamarnik and D. Katz. On deciding stability of multiclass queueing networks under buffer priority scheduling policies. *Annals of Applied Probability*, 19(5):2008–2037, 2009.

[14] M. Larrañaga, U. Ayesta, and I.M. Verloop. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation*, 70(10):841–858, 2013.

[15] T. Maertens, H. Bruneel, and J. Walraevens. Effect of class clustering on delay differentiation in priority scheduling. *Electronic Letters*, 48(10):568–569, 2012.

[16] 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 *Proceedings of the 7th International Conference on Queueing Theory and Network Applications (QTNA 7)*, Kyoto, Japan, 2012.

[17] H. Takagi. *Queueing analysis - vol. 3: discrete-time systems.* North Holland, 1993.

[18] O.S. Uluscu and T. Altiok. Waiting time approximation in multi-class queueing systems with multiple types of class-dependent interruptions. *Annals of Operations Research*, 202(1):185–195, 2013.

[19] I.M. 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.