Performance analysis of a discrete-time two-class global-FCFS queue with two servers and geometric service times

Herwig Bruneel[1], Willem Mélange[1], Joris Walraevens[1]
[1]SMACS Research Group
Department of Telecommunications and Information Processing
Stijn De Vuyst[2], Dieter Claeys[1,2]
[2]Department of Industrial Systems Engineering and Product Design
Ghent University - UGent (Belgium)

Abstract

In this paper, we analyze a discrete-time queueing model with two types (classes) of customers and two servers, one for each customer class. Although each server can only process one type of customers, all customers are accommodated in one single queue and served in their order of arrival, irrespective of their types. The numbers of customers arriving in the system from time slot to time slot are independent, but the types of consecutive customers are not necessarily independent. Specifically, we assume that a first-order Markovian correlation ("interclass correlation") exists between the types of subsequent customers in the arrival stream. The fact that multiple customers of the same type may arrive back-to-back and customers have to be served in their order of arrival, causes occasional under-utilization of the service capacity of the system, because some customers may not be able to reach their server owing to the presence of customers of the opposite type in front of them.

In this paper, we assume that the service times of both types of customers are independent, geometrically distributed random variables. The paper extends earlier work where all the service times were assumed to be of fixed length, either equal to 1 slot each, or equal to multiple slots. The fact that, in the present paper, service times are of variable length, entails that customers being served simultaneously can overtake each other, thus disturbing the original arrival order. This phenomenon did not occur in previous studies with fixed-length service times, and represents the main new element of the paper. It also complicates the analysis of the system considerably. Nevertheless, we are able to derive explicit expressions for the probability generating functions and the mean values of the main performance measures of the system, in terms of the original system parameters and one root of a non-linear equation. Our results reveal the impact of the interclass correlation and the variable nature of the service times on the achievable throughput, the (mean) number of customers in the system, the (mean) customer sojourn times, the (mean) unfinished work in the system, and related quantities.

Key words: queueing; discrete time; multi-class; interclass correlation; dedicated servers; global FCFS; geometric service times

1 Introduction and mathematical model

Classical *multi-class queueing models* deal with situations where multiple types (or classes) of customers compete for the use of the same resources; see, e.g., [11, 2, 19, 16, 27, 25, 1, 3] for some recent examples in various application areas. Usually, the resources to be shared are the facilities that are able to deliver the requested services to the customers, or, in queueing language, the "servers" of the queueing system. Very often, the different customer classes are characterized by either their distinct arrival characteristics, their different service requirements, their loss priorities ([28]), their service priorities ([21, 23, 26]), etc. Arriving customers are either accommodated in separate class-specific queues or in one global shared queue, but in general they require some kind of processing from the *same servers*. It is mainly this particular circumstance that causes the interdependence between the queueing performance of the individual customer classes.

In the present paper, we study a multi-class queueing model which is different from the above setting, in that each customer class now has its own dedicated server, i.e., customer classes do not compete for the same servers, but where the access to the servers is to be shared among different customer classes, i.e., arriving customers are accommodated in one common waiting line and can only reach the service area of the system in their order of arrival, regardless of the class they belong to. In this setting, it is the sharing of the same storage capacity and the strict adherence to the first-come-first-served (FCFS) service discipline that causes interaction between the various customer classes. Some obvious practical applications of this kind of model occur in the context of road networks (see, e.g., [22, 30, 29, 10]), when cars having different destinations use the same road section in front of a traffic junction, or input queues in the context of packet switches in the nodes of communication networks (see, e.g. [24, 4]), when information packets destined to different downstream nodes are stored in shared buffers.

Specifically, this paper considers an infinite-capacity discrete-time queueing model with two types (classes) of customers, each having their own dedicated server. Customer classes are named 1 and 2, servers are called A and B. Server A can only process customers of type 1 and server B can only deal with customers of type 2. Customers are served in their order of arrival, regardless of the class they belong to. In the context of single-server models, some authors have named this kind of service discipline "multi-class FIFO" (see, e.g. [17, 15, 11]), where, of course, FIFO stands for first-in-first-out. Here, because of the presence of two servers instead of just one, and for the sake of consistency with our earlier papers, we will label this service discipline "global first-come-first-served" ("global FCFS").

As in all discrete-time models, the time axis is divided into fixed-length intervals, referred to as *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. Arrivals in the system are assumed to occur independently from slot to slot. Their numbers are characterized by the probability mass function (pmf) $e(n)$ and the probability generating function (pgf) $E(z)$, i.e.,

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

$$E(z) \triangleq \sum_{n=0}^{\infty} e(n)\, z^n \ . \tag{1}$$

The mean arrival rate (per slot) is given by

$$\lambda = E'(1) \ . \tag{2}$$

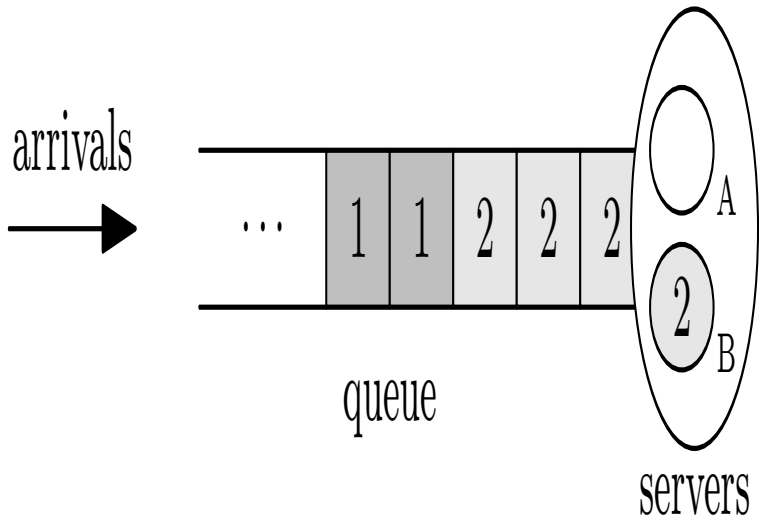

Figure 1: Situation where the system contains multiple customers and the eldest two customers belong to the same class.

Subsequent customers arriving in the system may belong to either class 1 or class 2 randomly, but not necessarily independently. Here we assume a simple first-order Markovian correlation between the types of consecutive customers; we refer to this correlation with the term "interclass correlation" in this paper. As in [9] and [8], we denote by α the probability that the next customer has *the same type as the previous one*, and by $1-\alpha$ the probability that the next customer belongs to *the opposite class as the previous one*. The parameter α ($0 \leq \alpha \leq 1$), referred to as the "cluster parameter" in the sequel, is indicative for the amount of class clustering in the arrival process: a high value of α implies that high numbers of customers belonging to the same class may arrive back-to-back, whereas a low value of α points at situations where the two customer types alternate more frequently. In general, the average length of a "cluster", i.e., a sequence of consecutive customers of the same type, either 1 or 2, is given by $1/(1-\alpha)$. In this model, both customer classes occur equally frequently in the long run, and thus represent half of the arrival load.

The service times of all customers, regardless of the class they belong to, are assumed to be geometrically distributed with parameter $1-\mu$, i.e. their pmf $s(n)$ is given by

$$s(n) \triangleq \text{Prob[service time} = n \text{ slots}] = \mu(1-\mu)^{n-1} \quad , \quad n \geq 1 \quad , \tag{3}$$

their pgf $S(z)$ is given by

$$S(z) = \frac{\mu z}{1-(1-\mu)z} \quad , \tag{4}$$

and their mean value $E[s]$ is

$$E[s] = \frac{1}{\mu} \quad . \tag{5}$$

We have studied a special case of the current model, where the service times are deterministically equal to 1 slot each, i.e., where $\mu = 1$, in our previous paper [9]. A first generalization of this was examined in [8], where arbitrary-length constant service times, i.e., service times equal to $s \geq 1$ slots each, were considered. This extension was motivated mainly by the desire to study the effect of the lengths of the service times (as compared to the slot length) on the queueing

behavior of the two-class global-FCFS system. Although seemingly simple, the extension proved to be non-trivial in terms of a substantially increased complexity of the state description of the system and its analysis. Mathematically speaking, the main reason for this is the fact that the deterministic distribution (with constant value s) does not possess the memoryless property if $s \neq 1$, which requires the state description of the system to include information on the elapsed (or remaining) service time(s) of the customer(s) in service. In the present paper, we want to tackle a second generalization of the basic model in [9]. Specifically, we want to relax the restriction that service times need to be deterministic and allow for variable-length service times, while still keeping the complexity of the analysis within reasonable limits. A geometric, and hence memoryless, distribution for the service times, is therefore a suitable choice. The variable nature of the service times implies that customers being served simultaneously (by servers A and B) can possibly "overtake" each other, i.e., a customer arriving later than an other customer may finish its service earlier, owing to a shorter service time. This phenomenon complicates the queueing analysis of the system considerably, because it disturbs the concepts of "previous customer" and "next customer" which are crucial in our Markovian model of class clustering. Nevertheless, we are able to derive explicit expressions for the pgfs of the system content, the server content, the queue content, and the unfinished work, in terms of the original system parameters and one root of a nonlinear equation. From these pgfs, we can derive, among others, the mean system content, the mean server content, the mean queue content, the mean unfinished work in the system, and the mean delay and mean waiting time of the customers.

It is worth mentioning that the current paper is also related with the work reported in reference [20], in which a continuous-time two-server queueing model with two types of customers and global FCFS service discipline is considered, under the assumptions that arrivals occur according to a classical continuous-time Poisson process and all service times are exponentially distributed. Although the continuous-time model and, especially, the method of analysis in [20] are profoundly different from the discrete-time approach presented in the current paper, both papers do examine the same type of queueing phenomenon. Some comparison of the results of both approaches will therefore also be briefly included further in this paper.

Specifically, the structure of the rest of this paper is as follows. In section 2, we subsequently introduce a Markovian state description of the system under study, establish equations describing the system-state evolution as a function of time, and determine the maximum allowable traffic intensity in order for the system to remain stochastically stable. Next, we present a detailed analysis of the system content and various other related performance measures. Section 3 treats two simple special cases of the model under study, for which results known in literature are easily retrieved. An extended discussion of the main results, including comparison with the results reported in [8, 20] and various numerical examples, is presented in section 4. Section 5 formulates some conclusions and briefly comments on possible future work.

2 Queueing analysis

2.1 Markovian state description of the system

Let u_k denote the total *system content*, i.e., the total number of customers present in the system (i.e., queue + servers, see Figs. 1 and 2), let q_k denote the *queue content*, i.e., the number of waiting customers (excluding those in service, if any) in the system, and let r_k $(0 \leq r_k \leq 2)$ denote the *server content*, i.e., the number of customers in service (servers A and B together), as

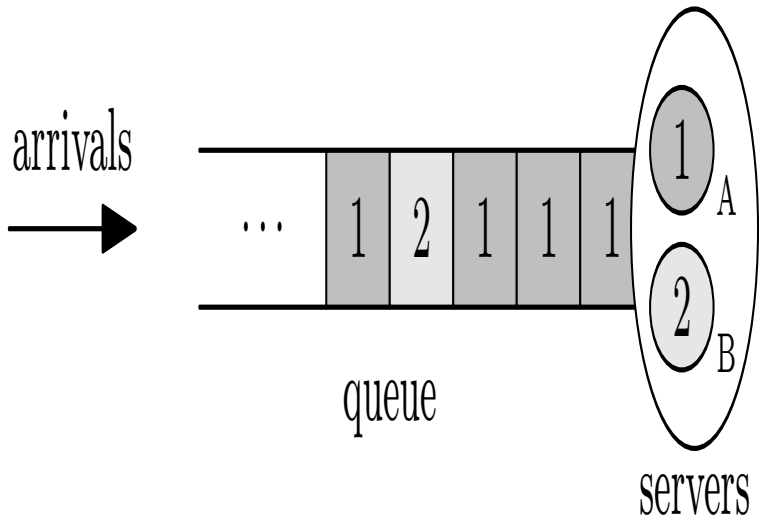

Figure 2: Situation where the system contains multiple customers and the eldest two customers belong to opposite classes.

observed at the beginning of the k-th slot. It is clear then that, for all values of k, the following simple relationship exists between these random variables:

$$u_k = q_k + r_k \quad . \tag{6}$$

Furthermore, we note that the possible values of the random variable r_k are determined by the value of the random variable u_k; specifically, we have

$$\begin{aligned} u_k = 0 &\Rightarrow r_k = 0 \quad , \\ u_k = 1 &\Rightarrow r_k = 1 \quad , \\ u_k > 1 &\Rightarrow r_k = 1 \text{ or } 2 \quad . \end{aligned} \tag{7}$$

Equations (7) can be justified as follows. If $u_k = 0$, then the system is empty and, consequently, the servers do not contain any customers either; so $r_k = 0$. If $u_k = 1$, then there is exactly one customer in the system, which occupies either server A or server B, and, hence, $r_k = 1$. If $u_k > 1$, then at least two customers are in the system. If the eldest two customers belong to the same class, then only one customer can be served and $r_k = 1$. This case is illustrated in Fig. 1 where the eldest two customers happen to be of type 2. If the eldest two customers belong to opposite classes, as illustrated in Fig. 2, then both are served and $r_k = 2$.

One of the main purposes of our analysis is to derive the steady-state distribution of the number of customers present in either the system or the queue. Therefore, it is clear that at least either the random variable u_k or the random variable q_k should be a component of the system state description at slot k. As new arrivals occur independently from slot to slot, there is no need to keep track of the arrivals in the system state. However, the state description should also contain sufficient information on the number of customers that can possibly leave the system at the end of each slot. One might be tempted to conclude that the state of the system at slot k can be fully described by the vector (u_k, r_k). In order for this to be true, the joint probability

distribution of the couple (u_{k+1}, r_{k+1}) should be completely determined by the value of the couple (u_k, r_k). It turns out that this is not really true, although one could say that it is nearly true. More specifically, we shall see that the distribution of u_{k+1} can be derived as soon as the value of (u_k, r_k) is known, but the distribution of r_{k+1} is only fully determined by the couple (u_k, r_k) if $u_k \neq 1$.

In order to further substantiate the above claims, we first introduce some further notation. Let ℓ_k indicate the number of customers leaving the system, i.e., the number of departures from the system, at the end of slot k. Then ℓ_k can be expressed as

$$\ell_k = \sum_{i=1}^{r_k} t_{ik} \ , \tag{8}$$

where the quantities $\{t_{ik}\}$ are independent and identically distributed Bernoulli random variables with parameter μ: $t_{ik} = 1$ if the i-th customer in service during slot k completes service at the end of slot k, whereas $t_{ik} = 0$ if the service continues after slot k; owing to the memoryless nature of the geometric service-time distribution, these events occur with probabilities μ and $1 - \mu$ respectively. It is not difficult to see then that

$$u_{k+1} = u_k + e_k - \ell_k = u_k + e_k - \sum_{i=1}^{r_k} t_{ik} \ , \tag{9}$$

where e_k, with pmf $e(n)$ and pgf $E(z)$ as defined in (1), denotes the total number of arrivals in the system during the k-th slot. As the distributions of e_k and $\{t_{ik}\}$ are known, equation (9) proves that the distribution of u_{k+1} is fully determined by the value of the couple (u_k, r_k).

Let us turn now to the distribution of the random variable r_{k+1}, the number of customers in service during slot $k+1$. If $u_k = 0$, then also $r_k = 0$, and $u_{k+1} = e_k$, which (in view of (7)) implies that r_{k+1} is completely determined by the number and types of the new arrivals in slot k. Since all previous customers have left the system already, the system takes a new start at the beginning of slot $k+1$ and no more information on the types of earlier customers needs to be retained. If $u_k > 1$, then either $r_k = 1$ or $r_k = 2$, but in both cases the last customer to have entered service is still present in the system, so that "the distribution of" the type of the next customer to enter service can be determined from the Markovian customer-class-correlation model. However, if $u_k = 1$, then the customer in service is either the last customer to have entered the system or a customer of the opposite type that has been overtaken by this last customer. In this case, the couple (u_k, r_k) does not contain all the necessary information on the system state at slot k to allow further study of the system state at slots $k+1$ and later. It is clear that in this case, the system state must be supplemented with (specific) information on the type of the (only) customer in the system.

Summarizing, we may conclude that the state of the queueing system at hand can be fully determined by means of a vector of three random variables: (u_k, v_k, r_k), where the first and the third components were defined earlier and where the second component v_k is defined as follows. If $u_k = 0$, then, by definition, $v_k = 0$. If $u_k = 1$, then, $v_k = 1$ if and only if the last customer that has entered the system is still in service, whereas $v_k = 0$ if and only if this condition does not hold (and, thus, the customer in service has the opposite type). Finally, if $u_k > 1$, then, again by definition, $v_k = 1$ if and only if $r_k = 1$, whereas $v_k = 0$ if and only if $r_k = 2$. Notice that the second component v_k of the state vector is actually only needed when $u_k = r_k = 1$. By also defining v_k in case $u_k \neq 1$, as we did above, we can reduce the state vector to just two components: (u_k, v_k),

the third component r_k being completely determined by this couple. Indeed, when $u_k = 0$ or $u_k = 1$, we know that $r_k = u_k$, and, when $u_k > 1$, we know that $r_k = 1$ if $v_k = 1$ and $r_k = 2$ if $v_k = 0$. In the sequel, we will therefore basically use the couple (u_k, v_k) as the state description at slot k, but also keep the notation r_k when useful.

2.2 System-state evolution

Let us now examine the evolution of the state vector from slot k to slot $k+1$. In order to do so, we assume that both components at slot k, i.e., u_k and v_k are known, and we try to determine (the joint distribution of) the state components at slot $k+1$, i.e., u_{k+1} and v_{k+1}, conditioned on this knowledge. As far as u_{k+1} is concerned, all the required information is already contained in equation (9) above:

$$u_{k+1} = u_k + e_k - \sum_{i=1}^{r_k} t_{ik} \quad , \tag{10}$$

because knowledge of (u_k, v_k) implies knowledge of (u_k, r_k) and the distributions of e_k and $\{t_{ik}\}$ are known. The virtue of equation (10) is that it expresses u_{k+1} explicitly in terms of the known random variables e_k and $\{t_{ik}\}$, as soon as the current state (u_k, v_k) is given. It is intuitively clear that the second component of the state vector at slot $k+1$, i.e., v_{k+1} is also dependent on the random variables e_k and $\{t_{ik}\}$, but, unfortunately, we do not dispose of a simple equation that expresses v_{k+1} in terms of these. Rather, we must distinguish between various values of the current system state (u_k, v_k) to be able to express the dependence of v_{k+1} on e_k and $\{t_{ik}\}$. Moreover, it also turns out to be impossible to treat the dependence of v_{k+1} on e_k and $\{t_{ik}\}$ separately from the dependence of u_{k+1} on the same variables. In view of all the above considerations, we have found that the most practical way to proceed is to introduce the conditional joint pgfs $P(z, x|n, j)$, defined as follows:

$$P(z, x|n, j) \triangleq E[z^{u_{k+1}} x^{v_{k+1}} | u_k = n, v_k = j] \quad , \tag{11}$$

for all $n \geq 0$ and $j \in \{0, 1\}$. Here the notation $E[\cdot]$ refers to the mean-value operator.

Let us first consider the case where $r_k = 0$, and, hence, also $u_k = 0$ and $v_k = 0$. Equation (10) shows that in this case, the system-state evolution (and, hence, also v_{k+1}) is completely determined by the new arrivals in slot k (i.e., the quantity e_k with pmf $e(n)$ as defined in (1)) and we find

$$P(z, x|0, 0) = E[z^{e_k} x^{v_{k+1}} | u_k = 0, v_k = 0] = \varphi(z, x) \quad , \tag{12}$$

where $\varphi(z, x)$ is a known function of z and x, defined as

$$\varphi(z, x) \triangleq e(0) + e(1)zx + [E(z) - e(0) - e(1)z](1 - \alpha + \alpha x) \quad . \tag{13}$$

Next, we turn to the case where $r_k = 1$. This situation may occur when the system contains only one single customer (i.e., $u_k = 1$, and either $v_k = 1$ or $v_k = 0$) or when the system contains at least two customers and the eldest two customers belong to the same class (i.e., $u_k > 1$ and $v_k = 1$). According to equation (10), state changes are now completely determined by the joint effect of the new arrivals in slot k (i.e., the quantity e_k with pmf $e(n)$) and the service of the customer being served (i.e., the Bernoulli quantity t_{1k}, with $\text{Prob}[t_{1k} = 0] = 1 - \mu$ and $\text{Prob}[t_{1k} = 1] = \mu$). We distinguish between four different sub-cases:

$$
\begin{aligned}
P(z,x|1,1) &= E\big[z^{1+e_k-t_{1k}}x^{v_{k+1}}|u_k=1,v_k=1\big] = (1-\mu)\psi(z,x)+\mu\varphi(z,x) \ , \\
P(z,x|1,0) &= E\big[z^{1+e_k-t_{1k}}x^{v_{k+1}}|u_k=1,v_k=0\big] = (1-\mu)\hat{\psi}(z,x)+\mu\varphi(z,x) \ , \\
P(z,x|2,1) &= E\big[z^{2+e_k-t_{1k}}x^{v_{k+1}}|u_k=2,v_k=1\big] = (1-\mu)z^2E(z)x+\mu\psi(z,x) \ , \\
P(z,x|n,1) &= E\big[z^{n+e_k-t_{1k}}x^{v_{k+1}}|u_k=n,v_k=1\big] = [(1-\mu)zx+\mu(1-\alpha+\alpha x)]z^{n-1}E(z) \ , \text{ if } n>2 \ .
\end{aligned}
\tag{14}
$$

Here the known functions $\psi(z,x)$ and $\hat{\psi}(z,x)$ are defined as

$$
\begin{aligned}
\psi(z,x) &\triangleq e(0)zx + z[E(z)-e(0)](1-\alpha+\alpha x) \ , \\
\hat{\psi}(z,x) &\triangleq e(0)z + z[E(z)-e(0)][\alpha+(1-\alpha)x] \ .
\end{aligned}
\tag{15}
$$

Finally, we examine the case where $r_k = 2$. This situation occurs when at least two customers are in the system and the eldest two customers have opposite types (i.e., $u_k > 1$ and $v_k = 0$). The system-state evolution is now completely determined by the joint effect of the new arrivals in slot k (i.e., the quantity e_k with pmf $e(n)$) and the services of the two customers being served (i.e., the Bernoulli quantities t_{1k} and t_{2k}). We make a further distinction between three different sub-cases:

$$
\begin{aligned}
P(z,x|2,0) &= E\big[z^{2+e_k-t_{1k}-t_{2k}}x^{v_{k+1}}|u_k=2,v_k=0\big] \\
&= (1-\mu)^2z^2E(z)+\mu(1-\mu)[\psi(z,x)+\hat{\psi}(z,x)]+\mu^2\varphi(z,x) \ , \\
P(z,x|3,0) &= E\big[z^{3+e_k-t_{1k}-t_{2k}}x^{v_{k+1}}|u_k=3,v_k=0\big] \\
&= (1-\mu)^2z^3E(z)+\mu(1-\mu)z^2E(z)(1+x)+\mu^2\psi(z,x) \ , \\
P(z,x|n,0) &= E\big[z^{n+e_k-t_{1k}-t_{2k}}x^{v_{k+1}}|u_k=n,v_k=0\big] \\
&= [(1-\mu)^2z^2+\mu(1-\mu)z(1+x)+\mu^2(1-\alpha+\alpha x)]z^{n-2}E(z) \ , \text{ if } n>3 \ .
\end{aligned}
\tag{16}
$$

2.3 Stability condition of the system

In the next subsections we will analyze the steady-state behavior of the queueing system under study. Before tackling this analysis, we first examine the conditions under which such a steady state exists. In general terms, it is not difficult to see that the system is stable, i.e., a steady state exists, if and only if the traffic intensity, i.e., the average amount of *work* entering the system per slot, given by $\lambda E[s] = \lambda/\mu$, is strictly less than the average "service capacity" of the system, i.e., the average amount of *work* that the servers are able to deliver per slot when the system is saturated, i.e., when there are always customers available in the system. As each busy server performs one unit of work per time slot, the amount of work the system-as-a-whole delivers in one slot, say slot k, is equal to the number of busy servers during that slot, i.e., the quantity r_k,

defined earlier. The average service capacity of the system is therefore given by

$$\begin{aligned} C &= \lim_{k\to\infty} E[r_k|u_k \text{ large}] \\ &= 1\cdot \lim_{k\to\infty} \text{Prob}[r_k=1|u_k \text{ large}] + 2\cdot \lim_{k\to\infty} \text{Prob}[r_k=2|u_k \text{ large}] \\ &= 1\cdot \lim_{k\to\infty} \text{Prob}[v_k=1|u_k \text{ large}] + 2\cdot \lim_{k\to\infty} \text{Prob}[v_k=0|u_k \text{ large}] \\ &= v_{\text{sat}}(1) + 2v_{\text{sat}}(0) \ , \end{aligned} \tag{17}$$

where the quantities $v_{\text{sat}}(1)$ and $v_{\text{sat}}(0)$ denote the long-run probabilities that the server state (i.e., v_k for $k\to\infty$) is either 1 or 0 in a saturated system. The probabilities $v_{\text{sat}}(1)$ and $v_{\text{sat}}(0)$ can be determined from a study of the time evolution of the server state in a saturated system. In order to do so, we use the last equation in the set (14) and the last equation in the set (16), for $z=1$ and large values of n, which results in

$$P(1,x|n,1) \triangleq E[x^{v_{k+1}}|u_k=n, v_k=1] = (1-\mu)x + \mu(1-\alpha+\alpha x) \ ,$$

$$P(1,x|n,0) \triangleq E[x^{v_{k+1}}|u_k=n, v_k=0] = (1-\mu)^2 + \mu(1-\mu)(1+x) + \mu^2(1-\alpha+\alpha x) \ .$$

In terms of conditional probabilities, the above equations imply that, in a saturated system,

$$\text{Prob}[v_{k+1}=1|v_k=1]_{\text{sat}} = 1-\mu+\mu\alpha \quad , \quad \text{Prob}[v_{k+1}=0|v_k=1]_{\text{sat}} = \mu(1-\alpha) \ ,$$

$$\text{Prob}[v_{k+1}=1|v_k=0]_{\text{sat}} = \mu(1-\mu+\mu\alpha) \quad , \quad \text{Prob}[v_{k+1}=0|v_k=0]_{\text{sat}} = 1-\mu+\mu^2(1-\alpha) \ .$$

It follows that the probabilities $v_{\text{sat}}(1)$ and $v_{\text{sat}}(0)$ satisfy the following set of linear balance equations:

$$\begin{aligned} v_{\text{sat}}(1) &= v_{\text{sat}}(1)(1-\mu+\mu\alpha) + v_{\text{sat}}(0)\mu(1-\mu+\mu\alpha) \ , \\ v_{\text{sat}}(0) &= v_{\text{sat}}(1)\mu(1-\alpha) + v_{\text{sat}}(0)[1-\mu+\mu^2(1-\alpha)] \ . \end{aligned} \tag{18}$$

Of course, $v_{\text{sat}}(1)$ and $v_{\text{sat}}(0)$ should also add up to 1, i.e.,

$$v_{\text{sat}}(1) + v_{\text{sat}}(0) = 1 \ . \tag{19}$$

Solving equations (18) and (19), we obtain

$$v_{\text{sat}}(1) = \frac{1-\mu+\mu\alpha}{1+(1-\mu)(1-\alpha)} \quad , \quad v_{\text{sat}}(0) = \frac{1-\alpha}{1+(1-\mu)(1-\alpha)} \ , \tag{20}$$

and, hence, from (17),

$$C = \frac{1+(2-\mu)(1-\alpha)}{1+(1-\mu)(1-\alpha)} \ , \tag{21}$$

so that the stability condition of the system can be expressed as

$$\frac{\lambda}{\mu} < \frac{1+(2-\mu)(1-\alpha)}{1+(1-\mu)(1-\alpha)} \ . \tag{22}$$

2.4 Steady-state analysis of the system content

For all k, let $p_k(n,j)$ and $P_k(z,x)$ denote the joint pmf and the joint pgf of the state vector variables (u_k, v_k), respectively, i.e.,

$$p_k(n,j) \triangleq \text{Prob}[u_k = n, v_k = j] \quad , \quad n \geq 0, \;\; j \in \{0,1\} \;\; ,$$

$$P_k(z,x) \triangleq E[z^{u_k} x^{v_k}] = \sum_{n=0}^{\infty} \sum_{j=0}^{1} p_k(n,j) z^n x^j \;\; .$$

Then, by virtue of the law of total expectation, equation (11) entails

$$P_{k+1}(z,x) = E[z^{u_{k+1}} x^{v_{k+1}}] = \sum_{n=0}^{\infty} \sum_{j=0}^{1} p_k(n,j) P(z,x|n,j) \;\; . \tag{23}$$

Now, let us assume that the queueing system at hand is stable, i.e., that the stability condition (22) is fulfilled. Letting the time parameter k go to infinity in equation (23) results in

$$P(z,x) = \sum_{n=0}^{\infty} \sum_{j=0}^{1} p(n,j) P(z,x|n,j) \;\; , \tag{24}$$

where

$$p(n,j) \triangleq \lim_{k \to \infty} p_k(n,j) \quad \text{and} \quad P(z,x) \triangleq \lim_{k \to \infty} P_k(z,x) \tag{25}$$

are the steady-state pmf and the steady-state pgf of the system state vector, respectively. By means of (12), (14) and (16), equation (24) can be rewritten as

$$\begin{aligned} P(z,x) = \; & p(0,0)\varphi(z,x) + p(1,1)[(1-\mu)\psi(z,x) + \mu\varphi(z,x)] \\ & + p(1,0)[(1-\mu)\hat{\psi}(z,x) + \mu\varphi(z,x)] + p(2,1)[(1-\mu)z^2 E(z)x + \mu\psi(z,x)] \\ & + \sum_{n=3}^{\infty} p(n,1)[(1-\mu)zx + \mu(1-\alpha+\alpha x)]z^{n-1}E(z) \\ & + p(2,0)[(1-\mu)^2 z^2 E(z) + \mu(1-\mu)[\psi(z,x) + \hat{\psi}(z,x)] + \mu^2\varphi(z,x)] \\ & + p(3,0)[(1-\mu)^2 z^3 E(z) + \mu(1-\mu)z^2 E(z)(1+x) + \mu^2\psi(z,x)] \\ & + \sum_{n=4}^{\infty} p(n,0)[(1-\mu)^2 z^2 + \mu(1-\mu)z(1+x) + \mu^2(1-\alpha+\alpha x)]z^{n-2}E(z) \;\; . \end{aligned} \tag{26}$$

Introducing the partial generating functions $U_1(z)$ and $U_0(z)$ as

$$U_1(z) \triangleq \sum_{n=1}^{\infty} p(n,1) z^n \quad \text{and} \quad U_0(z) \triangleq \sum_{n=0}^{\infty} p(n,0) z^n \;\; , \tag{27}$$

and using the expressions (13) and (15), we can rewrite the above equation as

$$
\begin{aligned}
U_1(z)x + U_0(z) = \;& p(0,0)[e(0)\alpha - e(1)(1-\alpha)z](1-x) + p(0,0)E(z)(1-\alpha+\alpha x) \\
& + p(1,1)\{e(0)[(1-\alpha)(1-\mu)z - \alpha\mu] + e(1)(1-\alpha)\mu z\}(x-1) \\
& + p(1,1)E(z)[(1-\mu)z + \mu](1-\alpha+\alpha x) \\
& + p(1,0)\{e(0)[(1-\alpha)(1-\mu)z + \alpha\mu] - e(1)(1-\alpha)\mu z\}(1-x) \\
& + p(1,0)E(z)[(1-\alpha)(1-\mu)zx + \alpha\mu x + (1-\mu)\alpha z + \mu(1-\alpha)]\} \\
& + p(2,1)\{e(0)(1-\alpha)\mu z(x-1) + E(z)[(1-\mu)z^2 x + \mu z(1-\alpha+\alpha x)]\} \\
& + p(2,0)\{[e(0)\alpha - e(1)(1-\alpha)z]\mu^2(1-x) + E(z)[(1-\mu)^2 z^2 + \mu(1-\mu)z(x+1) + \mu^2(1-\alpha+\alpha x)]\} \\
& + p(3,0)\{e(0)(1-\alpha)\mu^2 z(x-1) + zE(z)[(1-\mu)^2 z^2 + \mu(1-\mu)z(x+1) + \mu^2(1-\alpha+\alpha x)]\} \\
& + \left[U_1(z) - \sum_{n=1}^{2} p(n,1)z^n\right]\frac{E(z)}{z}[(1-\mu)zx + \mu(1-\alpha+\alpha x)] \\
& + \left[U_0(z) - \sum_{n=0}^{3} p(n,0)z^n\right]\frac{E(z)}{z^2}[(1-\mu)^2 z^2 + \mu(1-\mu)z(1+x) + \mu^2(1-\alpha+\alpha x)] \;\; .
\end{aligned}
\tag{28}
$$

By grouping, in the above equation, terms containing $p(2,1)$, $p(2,0)$ and $p(3,0)$, we get

$$
\begin{aligned}
U_1(z)x + U_0(z) = \;& p(0,0)[e(0)\alpha - e(1)(1-\alpha)z](1-x) + p(0,0)E(z)(1-\alpha+\alpha x) \\
& + p(1,1)\{e(0)[(1-\alpha)(1-\mu)z - \alpha\mu] + e(1)(1-\alpha)\mu z\}(x-1) \\
& + p(1,1)E(z)[(1-\mu)z + \mu](1-\alpha+\alpha x) \\
& + p(1,0)\{e(0)[(1-\alpha)(1-\mu)z + \alpha\mu] - e(1)(1-\alpha)\mu z\}(1-x) \\
& + p(1,0)E(z)[(1-\alpha)(1-\mu)zx + \alpha\mu x + (1-\mu)\alpha z + \mu(1-\alpha)] \\
& + p(2,1)e(0)(1-\alpha)\mu z(x-1) + p(2,0)[e(0)\alpha - e(1)(1-\alpha)z]\mu^2(1-x) + p(3,0)e(0)(1-\alpha)\mu^2 z(x-1) \\
& + \left[U_1(z) - p(1,1)z\right]\frac{E(z)}{z}[(1-\mu)zx + \mu(1-\alpha+\alpha x)] \\
& + \left[U_0(z) - p(0,0) - p(1,0)z\right]\frac{E(z)}{z^2}[(1-\mu)^2 z^2 + \mu(1-\mu)z(1+x) + \mu^2(1-\alpha+\alpha x)] \;\; .
\end{aligned}
\tag{29}
$$

Apart from the two partial pgfs $U_1(z)$ and $U_0(z)$, equation (29) contains six unknown probabilities: $p(0,0)$, $p(1,1)$, $p(1,0)$, $p(2,1)$, $p(2,0)$, and $p(3,0)$. It turns out that three simple linear

relationships can be established between these six unknowns by identifying the constant terms and the linear terms in the variables x and z on both sides of equation (29). Specifically, for the terms in z^0x^0, we obtain

$$p(0,0) = p(0,0)e(0) + p(1,1)\mu e(0) + p(1,0)\mu e(0) + p(2,0)\mu^2 e(0) \ , \tag{30}$$

for the terms in z^1x^1,

$$\begin{aligned} p(1,1) = \ & p(0,0)e(1) + p(1,1)[(1-\mu)e(0) + \mu e(1)] + p(1,0)\mu e(1) + p(2,1)\mu e(0) \\ & + p(2,0)\mu^2 e(1) + p(2,0)\mu(1-\mu)e(0) + p(3,0)\mu^2 e(0) \ . \end{aligned} \tag{31}$$

and for the terms in z^1x^0,

$$p(1,0) = p(1,0)(1-\mu)e(0) + p(2,0)\mu(1-\mu)e(0) \ . \tag{32}$$

Note that the three "boundary equations" (30), (31) and (32) can also be derived by expressing the probabilities of finding the system in states $(0,0)$, $(1,1)$ and $(1,0)$, respectively, in terms of the state probabilities one slot earlier, once a steady state has been reached. The three boundary equations can now be used to eliminate the unknown probabilities $p(2,1)$, $p(2,0)$ and $p(3,0)$ from the functional equation (29), which results in

$$\begin{aligned} U_1(z)x + U_0(z) = \ & p(0,0)[\alpha(1-x) + E(z)(1-\alpha+\alpha x)] + p(1,1)(1-\alpha)z[(1-\mu)E(z)-1](1-x) \\ & + p(1,0)E(z)[(1-\mu)(1-\alpha)zx + \mu\alpha x + (1-\mu)\alpha z + \mu(1-\alpha)] + p(1,0)(1-\alpha)z(1-x) \\ & + U_1(z)\frac{E(z)}{z}[(1-\mu)zx + \mu(1-\alpha+\alpha x)] \\ & + \left[U_0(z) - p(0,0) - p(1,0)z\right]\frac{E(z)}{z^2}[(1-\mu)^2z^2 + \mu(1-\mu)z(1+x) + \mu^2(1-\alpha+\alpha x)] \ . \end{aligned} \tag{33}$$

By identifying equal powers of x on both sides of the above equation, we obtain a system of two linear equations in the partial pgfs $U_1(z)$ and $U_0(z)$:

$$\begin{aligned} U_1(z) = \ & p(0,0)\alpha[E(z)-1] - p(1,1)(1-\alpha)z[(1-\mu)E(z)-1] \\ & + p(1,0)\{\mu\alpha E(z) + (1-\alpha)z[(1-\mu)E(z)-1]\} + U_1(z)\frac{E(z)}{z}[(1-\mu)z + \mu\alpha] \\ & + \left[U_0(z) - p(0,0) - p(1,0)z\right]\frac{E(z)}{z^2}\mu[(1-\mu)z + \mu\alpha] \end{aligned} \tag{34}$$

and

$$\begin{aligned} U_0(z) = \ & p(0,0)[\alpha + (1-\alpha)E(z)] + p(1,1)(1-\alpha)z[(1-\mu)E(z)-1] \\ & + p(1,0)\{z + \mu(1-\alpha)E(z) + \alpha z[(1-\mu)E(z)-1]\} + U_1(z)\frac{E(z)}{z}\mu(1-\alpha) \\ & + \left[U_0(z) - p(0,0) - p(1,0)z\right]\frac{E(z)}{z^2}[(1-\mu)^2z^2 + \mu(1-\mu)z + \mu^2(1-\alpha)] \ . \end{aligned} \tag{35}$$

Solving the two simultaneous linear equations (34) and (35) for the two partial pgfs $U_1(z)$ and $U_0(z)$ is a matter of some standard, but tedious, algebraic operations. Once the solution of the system (34)(35) is known, the steady-state pgf $U(z)$ of the system content can be obtained as

$$U(z) = P(z,1) = U_1(z) + U_0(z) \ , \tag{36}$$

where the function $P(z,x)$ was defined in (25). It turns out that the result can be written as

$$U(z) = \mu(z-1)E(z)\frac{p(0)g_0(z) + p(1)g_1(z)}{n(z)} \ , \tag{37}$$

where $p(0)$ and $p(1)$ are unknown coefficients, defined as

$$p(0) \triangleq p(0,0) \quad \text{and} \quad p(1) \triangleq \alpha p(1,0) + (1-\alpha)p(1,1) \tag{38}$$

and $g_0(z)$, $g_1(z)$ and $n(z)$ are known functions of z, given by

$$\begin{aligned} g_0(z) &\triangleq \mu(1-2\alpha+\alpha z) + [\mu\alpha + (2-\mu-\alpha)z][1-(1-\mu)E(z)] \ , \\ g_1(z) &\triangleq z[(1-\mu)z+\mu][1-(1-\mu)E(z)] \ , \\ n(z) &\triangleq \big\{z - [(1-\mu)z+\mu\alpha]E(z)\big\}\big\{z-(1-\mu)[(1-\mu)z+\mu]E(z)\big\} - \mu^2(1-\alpha)E(z) \ . \end{aligned} \tag{39}$$

It is remarkable that equation (37) contains only two unknown parameters (notably, $p(0)$ and $p(1)$) while the original system of simultaneous equations (34) and (35) contains three (notably, $p(0,0)$, $p(1,0)$ and $p(1,1)$). Apparently, in the determination of $U(z)$, the knowledge of the boundary probabilities $p(1,0)$ and $p(1,1)$ individually is not required, but only the combination $p(1)$ of these, as defined in (38). We note that a probabilistic interpretation can be given for the parameter $p(1)$: it denotes the joint probability that the system contains exactly one customer and the next customer to enter service has the opposite type as the customer in service.

It now remains for us to determine the two remaining unknowns $p(0)$ and $p(1)$. A first relation between $p(0)$ and $p(1)$ can be obtained from the normalization condition of the system-content distribution, i.e., the condition $U(1) = 1$. Some algebra on equation (37) leads to

$$p(1) = [1-p(0)][1+(2-\mu)(1-\alpha)] - \frac{\lambda}{\mu}[1+(1-\mu)(1-\alpha)] \ . \tag{40}$$

Using this result in (37), we can express $U(z)$ as

$$U(z) = (z-1)E(z)\frac{p(0)h_0(z) + h_1(z)}{n(z)} \ , \tag{41}$$

where $h_0(z)$ and $h_1(z)$ are given by

$$\begin{aligned} h_0(z) &\triangleq \mu\big\{g_0(z) - [1+(2-\mu)(1-\alpha)]g_1(z)\big\} \ , \\ h_1(z) &\triangleq \big\{\mu(1-\alpha) + (\mu-\lambda)[1+(1-\mu)(1-\alpha)]\big\}g_1(z) \ . \end{aligned} \tag{42}$$

Equation (41) contains just one remaining unknown parameter $p(0)$, which denotes the probability that the system be empty. In order to determine $p(0)$, we invoke the analyticity of the pgf $U(z)$.

Specifically, it can be formally shown that, as soon as the stability condition (22) is met, the denominator $n(z)$ of equation (41) has exactly two zeroes inside the closed unit disk of the complex z-plane: $z = 1$ and, say, $z = \theta < 1$. A proof of this statement, based on the methodology developed by Gail et al. [13], is given in the Appendix at the end of this paper. Both the zeroes $z = 1$ and $z = \theta$ must also be zeroes of the numerator of (41) since, in these circumstances, $U(z)$ is a regular pgf and remains bounded inside the closed unit disk. For the zero $z = 1$, this condition is clearly fulfilled in view of the factor $z-1$ in the numerator of (41); for the zero $z = \theta$, this condition leads to

$$p(0)h_0(\theta) + h_1(\theta) = 0 \ ,$$

and, hence,

$$p(0) = -\frac{h_1(\theta)}{h_0(\theta)} \ , \tag{43}$$

and, finally,

$$U(z) = (z-1)E(z)\frac{h_0(\theta)h_1(z) - h_1(\theta)h_0(z)}{h_0(\theta)n(z)} \ . \tag{44}$$

In equations (41) and (44) all quantities are known: the functions $h_0(z)$, $h_1(z)$ and $n(z)$ were defined in terms of the system parameters in (42) and (39) above, whereas the quantity θ is uniquely defined as the only root strictly inside the complex unit disk of the equation $n(z) = 0$, and the probability $p(0)$ follows from (43). Note that, in general, the value of θ is to be determined numerically.

The mean system content $E[u]$ can be found from equation (41) as

$$E[u] = U'(1) = \frac{E''(1)}{2(C\mu - \lambda)} + \frac{\lambda(1-\lambda)(C-1)}{(1-\alpha)(C\mu-\lambda)} + \frac{\{\mu^2(2-\mu)^2[1-p(0)] + \lambda(1-\mu)[\mu p(0) + \lambda(1-\mu) - \mu(3-\mu)]\}(C-1)}{\mu(C\mu-\lambda)} \ , \tag{45}$$

where the average service capacity C was defined in (21). Higher-order moments of the system content can be similarly derived by computing higher-order derivatives of the pgf $U(z)$.

2.5 Other performance measures

In subsection 2.4, we have determined the steady-state characteristics (pgf, mean value) of the *system content*. Various other performance measures can be derived from these results by means of some additional mathematical manipulations.

Let us first concentrate on the *unfinished work* in the system, i.e., the total amount of service time required to process all the customers in the system, at the beginning of an arbitrary time slot in steady state. Specifically, let f_k denote the unfinished work at the beginning of slot k. Then it is easily seen that

$$f_k = \sum_{j=1}^{q_k} s_j + \sum_{i=1}^{r_k} \hat{s}_i \ , \tag{46}$$

where the quantities s_j $(1 \le j \le q_k)$ denote the (total) service times of the q_k customers present in the queue at the beginning of slot k, and the random variables $\hat{s}_i$ $(1 \le i \le r_k)$ indicate the remaining service times of the r_k customers in service, at the same epoch. From equation (3) we

know that the s_j's are independent geometric random variables with parameter $1-\mu$. In view of the memoryless property of the geometric distribution, the same goes for the $\hat{s}_i$'s, which implies that f_k is, in fact, the sum of $q_k + r_k = u_k$ (see equation (6)) independent geometric random variables with parameter $1-\mu$:

$$f_k = \sum_{j=1}^{u_k} \tilde{s}_j \ , \tag{47}$$

where all the $\tilde{s}_j$'s are i.i.d. It then easily follows that the steady-state pgf of the unfinished work at the beginning of a slot is given by

$$F(z) \triangleq \lim_{k\to\infty} E\big[z^{f_k}\big] = U(S(z)) \ , \tag{48}$$

where $U(z)$ is given by (37) or (41) and $S(z)$ is the known pgf of the service times of the customers (see equation (4)). The mean unfinished work $E[f]$ can be easily derived from this as

$$E[f] = F'(1) = U'(1)S'(1) = E[u]\,E[s] = \frac{E[u]}{\mu} \ . \tag{49}$$

The steady-state distribution of the *server content* can also be easily determined. Let $r(n)$ denote the corresponding pmf, i.e.,

$$r(n) \triangleq \lim_{k\to\infty} \text{Prob}[r_k = n] \ , \tag{50}$$

then it is easily seen that

$$\begin{aligned} r(0) &= p(0,0) \ , \\ r(1) &= p(1,0) + p(1,1) + \sum_{n=2}^{\infty} p(n,1) = p(1,0) + U_1(1) \ , \\ r(2) &= \sum_{n=2}^{\infty} p(n,0) = U_0(1) - p(0,0) - p(1,0) \ . \end{aligned} \tag{51}$$

Here the partial pgfs $U_1(z)$ and $U_0(z)$ were defined in (27) and the quantities $U_1(1)$ and $U_0(1)$ can be readily derived by solving the set of simultaneous equations (34) and (35) for $z = 1$; some standard algebra leads to

$$\begin{aligned} U_1(1) &= 2 - 2p(0,0) - p(1,0) - \frac{\lambda}{\mu} \ , \\ U_0(1) &= 2p(0,0) + p(1,0) - 1 + \frac{\lambda}{\mu} \ . \end{aligned} \tag{52}$$

Using these results and (38) in (51), we get

$$\begin{aligned} r(0) &= p(0) \ , \\ r(1) &= 2[1 - p(0)] - \frac{\lambda}{\mu} \ , \\ r(2) &= \frac{\lambda}{\mu} - [1 - p(0)] \ . \end{aligned} \tag{53}$$

The steady-state pgf $R(z)$ of the server content follows from this as

$$R(z) = \sum_{n=0}^{2} r(n) z^n = 1 + \frac{\lambda}{\mu} z(z-1) - [1 - p(0)](z-1)^2 \ . \tag{54}$$

The mean server content $E[r]$ is given by

$$E[r] = r(1) + 2r(2) = R'(1) = \frac{\lambda}{\mu} \ . \tag{55}$$

The (simple) result in equation (55) can also be found without solving any balance equations, by merely expressing that, in the steady-state, the average amount of work leaving the system per slot, i.e., the average number of busy servers $E[r]$, should be in balance with the average amount of work entering the system per slot, i.e., the traffic intensity $\lambda E[s] = \lambda/\mu$. In particular, we note that although the knowledge of the empty-system probability $p(0)$ is required to determine the pmf (53) and the pgf (54) of the server content, the parameter $p(0)$ does not appear in expression (55) for the mean server content.

The steady-state pgf $Q(z)$ of the *queue content* can be derived from equation (6) as follows:

$$Q(z) = \lim_{k \to \infty} E[z^{q_k}] = \lim_{k \to \infty} E\left[z^{u_k - r_k}\right] \ . \tag{56}$$

Using the law of total expectation, we can rewrite this as

$$\begin{aligned} Q(z) &= p(0,0) + [p(1,0) + p(1,1)] + \sum_{n=2}^{\infty} p(n,1) z^{n-1} + \sum_{n=2}^{\infty} p(n,0) z^{n-2} \\ &= p(0,0) + p(1,0) + z^{-1} U_1(z) + z^{-2}[U_0(z) - p(0,0) - p(1,0) z] \\ &= p(0,0)[1 - z^{-2}] + p(1,0)[1 - z^{-1}] + z^{-1} U_1(z) + z^{-2} U_0(z) \\ &= \frac{p(0,0)(z^2 - 1) + p(1,0) z(z-1) + P(z,z)}{z^2} \ . \end{aligned} \tag{57}$$

Here the joint pgf $P(z,x) = U_0(z) + x U_1(z)$ was introduced in (25). The mean queue content $E[q]$ can be expressed as

$$E[q] = Q'(1) = 2p(0,0) + p(1,0) - 2 + P'(1,1) \ , \tag{58}$$

where the total derivative $P'(1,1)$ is given by

$$P'(1,1) = \frac{d}{dz}[U_0(z) + z U_1(z)]|_{z=1} = U_0'(1) + U_1'(1) + U_1(1) \ , \tag{59}$$

so that, in view of (36) and (52),

$$E[q] = E[u] - \frac{\lambda}{\mu} = E[u] - E[r] \ , \tag{60}$$

in full agreement with equations (6) and (55).

By applying (the discrete-time version of) Little's theorem [18, 7, 12] and/or using (5), we can also find the mean delay (system time) $E[d]$ and the mean waiting time $E[w]$ of a customer as

$$E[d] = \frac{E[u]}{\lambda} \tag{61}$$

and

$$E[w] = \frac{E[q]}{\lambda} = E[d] - E[s] = \frac{E[u]}{\lambda} - \frac{1}{\mu} \ . \tag{62}$$

3 Special cases

3.1 The case $\mu = 1$

In an earlier paper [9], we have studied the special case where $\mu = 1$, i.e., where the service times of the customers are deterministically equal to 1 slot each, by means of a simpler analysis. It is not immediately obvious from the formulas derived above that we do obtain the same results as in [9]. Let us check the main results. First of all, we note that, for $\mu = 1$, the average service capacity of the system, given in equation (21), reduces to

$$C = 2 - \alpha \ , \tag{63}$$

so that the stability condition (22) reads

$$\lambda < 2 - \alpha \ . \tag{64}$$

Furthermore, for $\mu = 1$, the probability $p(1,0)$ reduces to zero, because in that case all service times have the same length so that customers cannot overtake each other, and therefore, v_k cannot be zero when $u_k = 1$. It follows that the probability $p(1)$, defined in (38), is equal to

$$p(1) = (1-\alpha)p(1,1) = (1-\alpha)u(1) \tag{65}$$

and the pgf $U(z)$, given in (37), reduces to

$$U(z) = \mu(z-1)E(z)\frac{u(0)[z+1-\alpha] + u(1)(1-\alpha)z}{z^2 - (1-\alpha+\alpha z)E(z)} \ , \tag{66}$$

where

$$\begin{aligned} u(0) &\triangleq \lim_{k\to\infty} \text{Prob}[u_k = 0] = p(0) \ , \\ u(1) &\triangleq \lim_{k\to\infty} \text{Prob}[u_k = 1] = p(1,1) \ . \end{aligned} \tag{67}$$

These are exactly the results obtained in [9].

3.2 The case $\alpha = 1$

When the cluster parameter α is equal to 1, all customers belong to the same class and the system under study basically degenerates to a single-class single-server queue with geometric service times. The stability condition (22) now requires

$$\lambda < \mu \ , \tag{68}$$

the probability $p(1)$, given by (38), reduces to zero and the pgf $U(z)$ (37) of the system content becomes

$$U(z) = \mu(z-1)E(z)\frac{p(0)\{\mu(z-1) + [\mu + (1-\mu)z][1-(1-\mu)E(z)]\}}{\{z - [(1-\mu)z+\mu]E(z)\}\{z-(1-\mu)[(1-\mu)z+\mu]E(z)\}} \quad , \tag{69}$$

or, upon cancellation of a common factor in numerator and denominator,

$$U(z) = \frac{p(0)\mu(z-1)E(z)}{z - [(1-\mu)z+\mu]E(z)} \quad . \tag{70}$$

This result is well-known from many books and papers on discrete-time queues; see e.g. [14, 5, 6, 7].

4 Discussion of results and numerical examples

Having derived, in section 2, expressions for the main performance measures, we are now in a position to discuss the qualitative behavior of the system, and illustrate it quantitatively by means of some numerical examples. We also compare the results obtained here for geometric service times with the corresponding results derived for fixed-length service times in [9, 8] and even with the ones obtained in [20] for a continuous-time model with exponential service times.

As the main issue of this paper is the degradation of the average service capacity of the system due to the imposed global-FCFS queueing discipline, we first focus on equation (21), which expresses the mean service capacity C in terms of the system parameters μ and α, i.e.,

$$C = \frac{1 + (2-\mu)(1-\alpha)}{1 + (1-\mu)(1-\alpha)} \quad . \tag{71}$$

Using equation (5), we can also express this as

$$C = C_{\text{geom}}(E[s]\,, \alpha) \triangleq \frac{(3-2\alpha)E[s] - (1-\alpha)}{(2-\alpha)E[s] - (1-\alpha)} \quad , \tag{72}$$

where $E[s]$ indicates the mean service time. Formula (72) is represented graphically in Figs. 3 and 4, where the service capacity $C_{\text{geom}}(E[s]\,, \alpha)$ is depicted versus α (for given values of $E[s]$) and versus $E[s]$ (for given values of α), respectively.

Expression (72) and Fig. 3 reveal that the average service capacity, i.e., the maximum achievable throughput of the system, expressed in work units per slot, is very directly determined by the interclass correlation in the arrival process as described by the cluster parameter α. Specifically, when α increases from 0 to 1, the service capacity goes down from

$$C_{\text{geom}}(E[s]\,, 0) = \frac{3E[s] - 1}{2E[s] - 1} \tag{73}$$

to

$$C_{\text{geom}}(E[s]\,, 1) = 1 \quad . \tag{74}$$

Of course, the intuitive explanation of this is that the system becomes more and more non-work-conserving, i.e., the fraction of time that only one of the two available servers is busy gets larger and larger, as the customers tend to cluster more according to their types.

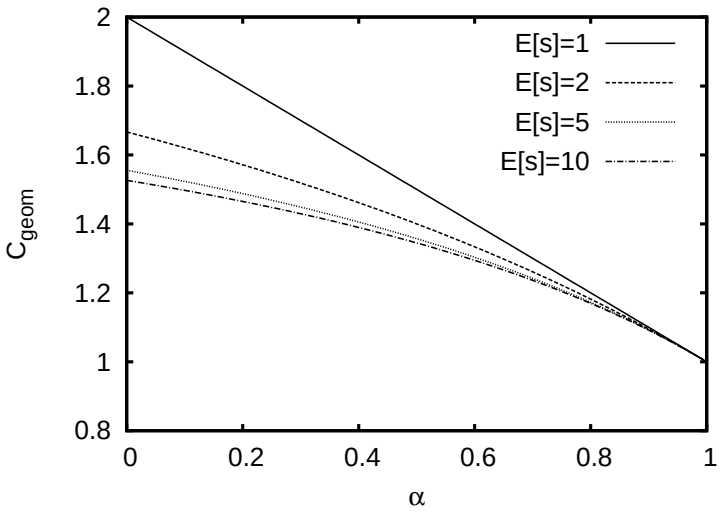

Figure 3: Maximum achievable throughput $C_{\rm geom}(E[s]\,,\alpha)$ for geometric service times, versus the cluster parameter α, for various values of the mean service time $E[s]$. For $E[s]=1$, the result is identical to the result for deterministic service times. For $E[s]\to\infty$, the result for exponential service times is obtained.

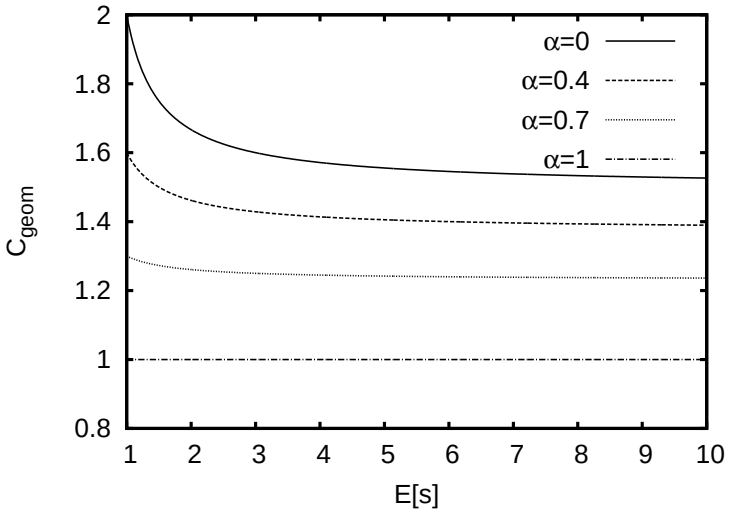

Figure 4: Maximum achievable throughput $C_{\rm geom}(E[s]\,,\alpha)$ for geometric service times, versus mean service time $E[s]$, for various values of the cluster parameter α.

Expression (72) and Fig. 4 also show that, for all values of $\alpha\neq 1$ (i.e., for genuine two-class systems), the maximum achievable throughput decreases when the average lengths of the service times get larger. This effect is more pronounced for low values of the cluster parameter α and gradually disappears as α approaches the value 1. Specifically, for a given value of $\alpha\neq 1$, the maximum achievable throughput decreases from

$$C_{\rm geom}(1,\alpha)=2-\alpha \tag{75}$$

to

$$C_{\rm geom}(\infty,\alpha)=\frac{3-2\alpha}{2-\alpha}\ , \tag{76}$$

when the mean service time goes from 1 to infinity. Fig. 4 illustrates, however, that the value of $C_{\rm geom}(E[s]\,,\alpha)$ saturates as soon as $E[s]$ has reached a threshold of about $3-5$ (depending on α). The reason for this (slight) decrease of service capacity for increasing service times, for low values of $E[s]$ and for any given value of $\alpha \neq 1$, is not immediately clear, in view of the observation that the fraction of customers that have the same type as the previous customer is simply given by the cluster parameter α and thus does not depend on $E[s]$. Of course, the phenomenon that customers of opposite types can overtake each other during service when $E[s] > 1$ plays a role here: even when customers of the two types alternate permanently (i.e., when $\alpha = 0$), the two servers only work simultaneously part of the time, because the order of arrival can be disturbed by the "overtaking mechanism". For $\alpha = 1$, of course, the system only processes one type of customers and the service capacity is constant and equal to 1 (see also equation (74)).

It is interesting to compare the above results for geometric service times with mean value $E[s]$, with the case of deterministic service times equal to $E[s]$ slots. From [8] we retrieve the following formula for the maximum achievable throughput for this case:

$$C_{\rm det}(E[s]\,,\alpha) = 2 - \alpha \ \ , \tag{77}$$

which apparently is independent of the mean service time. Specifically, we note that

$$C_{\rm det}(E[s]\,,\alpha) = C_{\rm geom}(1,\alpha) \ \ , \tag{78}$$

for all possible values of the mean service time $E[s]$. Graphically, this implies that the curve for geometric service times with $E[s] = 1$, i.e., the upper curve in Fig. 3, also gives the results for deterministic service times of any length. We recall from [8] that, in case of deterministic service times, the mean service capacity reduces by a factor 2 when the cluster parameter α goes up from the value 0 to the value 1, i.e.,

$$C_{\rm det}(E[s]\,,0) = 2 \tag{79}$$

and

$$C_{\rm det}(E[s]\,,1) = 1 \ \ . \tag{80}$$

Comparing (72) and (77), we can show algebraically that

$$C_{\rm geom}(E[s]\,,\alpha) \leq C_{\rm det}(E[s]\,,\alpha) \ \ , \tag{81}$$

for all values of $E[s]$ and α, i.e., the mean service capacity is lower for geometric service times than for deterministic service times with the same (mean) length. Of course, the inequality (81) is also very apparent from Fig. 3, where the curve for $E[s] = 1$ lies above all the other curves. Intuitively, again, this is due to the "overtaking mechanism" in case of geometric service times, which does not occur when all service times are identical.

We now turn to a comparison of the discrete-time model analyzed in the current paper with the continuous-time model treated in [20], which can be more or less considered as the continuous-time analog of the model under study. Specifically, the model in [20] assumes a continuous-time Poisson arrival process of new customers, with arrival rate λ (i.e., the inter-arrival times are exponential continuous random variables with parameter λ), two exponential servers with identical service rates μ (i.e., the service times are exponential continuous random variables with parameter μ) and an identical interclass correlation model with cluster parameter α as in the present paper. We note that, in view of the independent-increment property of the classical Poisson process (see, e.g., [18]), these assumptions imply in particular that the numbers of arrivals during non-overlapping time intervals are independent, just as in the current discrete-time model. Although the meaning

of the parameters λ and μ is different in the discrete time setting (where they refer to numbers of customers arriving or being served per finite time slot) and in the continuous time setting (where time slots as such do not exist and λ and μ refer to numbers of customers per time unit), the ratio λ/μ indicates the traffic intensity in both time settings, which in steady state is equal to the average number of busy servers, i.e., the mean server content, labelled $E[r]$ earlier in this paper (see, e.g., equation (55)). In reference [20], it was shown that the stability condition of the continuous-time system takes the form

$$\frac{\lambda}{\mu} < C_{\text{expon}}(\alpha) \triangleq \frac{3-2\alpha}{2-\alpha} \ . \tag{82}$$

This result is, in fact, consistent with our current findings, if we look at the continuous-time model as the limit of the discrete-time model when the slot length goes to zero. Indeed, in this limit transition, the parameters λ and μ in the discrete-time model would both go to zero, but their ratio would remain finite and the mean service time, expressed in infinitesimal-length time slots, would go to infinity, i.e., $E[s] \to \infty$. Comparing (82) with (72) or (76), we indeed find that

$$C_{\text{expon}}(\alpha) = C_{\text{geom}}(\infty, \alpha) \ . \tag{83}$$

It is not difficult to see from equations (72) and (82) that

$$C_{\text{expon}}(\alpha) \leq C_{\text{geom}}(E[s]\,, \alpha) \ , \tag{84}$$

for all values of $E[s]$ and α, i.e., the mean service capacity is lower for exponential service times than for geometric service times with the same (mean) length. The inequality (84) is also very clear from Fig. 3, where the curve for $E[s] \to \infty$ (nearly coinciding with the curve for $E[s] = 10$) lies below all the other curves.

We conclude from the above discussion, that for any given values of the parameters α and $E[s]$,

$$C_{\text{expon}}(\alpha) \leq C_{\text{geom}}(E[s]\,, \alpha) \leq C_{\text{det}}(E[s]\,, \alpha) \ , \tag{85}$$

which says that the case of geometric service times considered in this paper exhibits a throughput performance between the cases of exponential and deterministic service times, studied in [20] and [8], respectively. If we notice that, for a given value $E[s]$ of the mean service time, the variance of the service time is given by 0 (zero) in the deterministic case, by $E[s]\,[E[s]-1]$ in the geometric case, and by $E[s]^2$ in the exponential case, this result suggests that the deterioration of the maximum achievable throughput increases with the service-time variability. Intuitively, this could be attributed to the growing impact of the "overtaking mechanism" when consecutive service times can differ more.

We end this section with some numerical results for the discrete-time model with geometric service times. The main performance metric we focus on is the mean system content $E[u]$, as given by formula (45). Other interesting performance measures, such as the mean unfinished work $E[f]$ (see equation (49)), the mean queue content $E[q]$ (see equation (60)), the mean customer delay $E[d]$ (see equation (61)) and the mean waiting time $E[w]$ (see equation (62)) are not explicitly considered here, because of their very close relationship with $E[u]$. Although the formulas derived in this paper are valid for any choice of the arrival pgf $E(z)$, we choose a Poisson distribution for the number of arrivals per slot, i.e., $E(z) = e^{\lambda(z-1)}$, because of its great practical applicability and, also, because in this case the equation $n(z) = 0$ which has to be solved numerically to find the root θ is a transcendental equation (see equation (39)) and, hence more "complicated" than for many other choices of $E(z)$. Even in this case, no numerical problems were encountered.

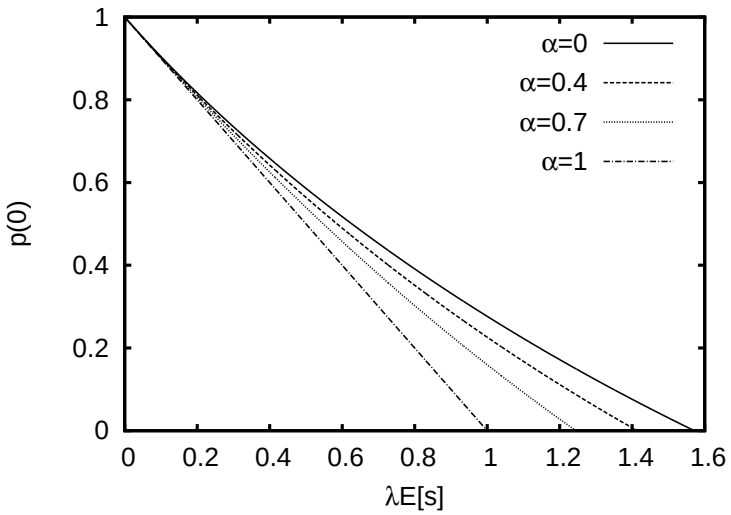

Figure 5: Probability of empty system $p(0)$ versus traffic intensity $\lambda E[s]$, for Poisson arrivals, geometric service times with mean $E[s] = 4$, and various values of the cluster parameter α.

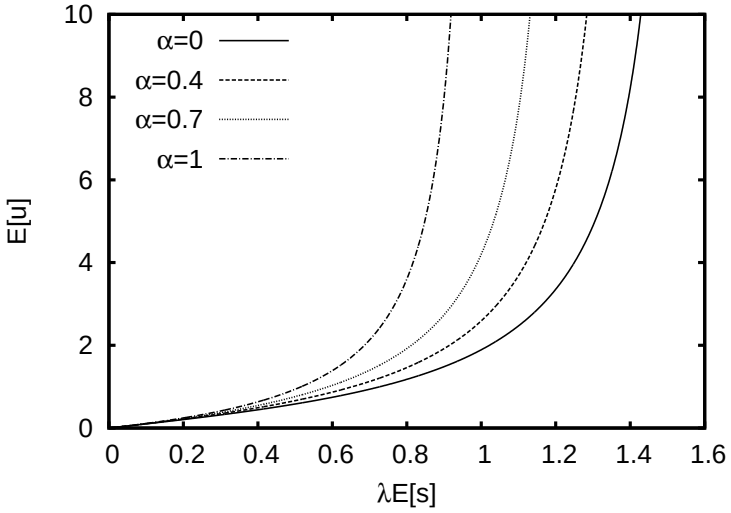

Figure 6: Mean system content $E[u]$ versus traffic intensity $\lambda E[s]$, for Poisson arrivals, geometric service times with mean $E[s] = 4$, and various values of the cluster parameter α.

Fig. 5 depicts the empty-system probability $p(0)$, determined numerically according to the procedure summarized in equation (43), versus the traffic intensity $\lambda E[s]$, for a value $E[s] = 4$. The figure clearly shows that the empty-system probability decreases as the traffic intensity gets larger, for any value of the cluster parameter α, as expected. The probability $p(0)$ reaches the value zero when the traffic intensity approaches the mean service capacity, i.e., when

$$\lambda E[s] = C_{\text{geom}}(4, \alpha) = \frac{11 - 7\alpha}{7 - 3\alpha} \quad . \tag{86}$$

For the α-values shown in Fig. 5, $p(0)$ reaches zero at traffic intensity 1.57 in case $\alpha = 0$, 1.41 in case $\alpha = 0.4$, 1.24 in case $\alpha = 0.7$,and 1 in case $\alpha = 1$.

Figs. 6 and 7 show graphs of the mean system content $E[u]$ versus the traffic intensity $\lambda E[s]$, for a given value $E[s] = 4$ and various values of α, and for a given value $\alpha = 0.5$ and various

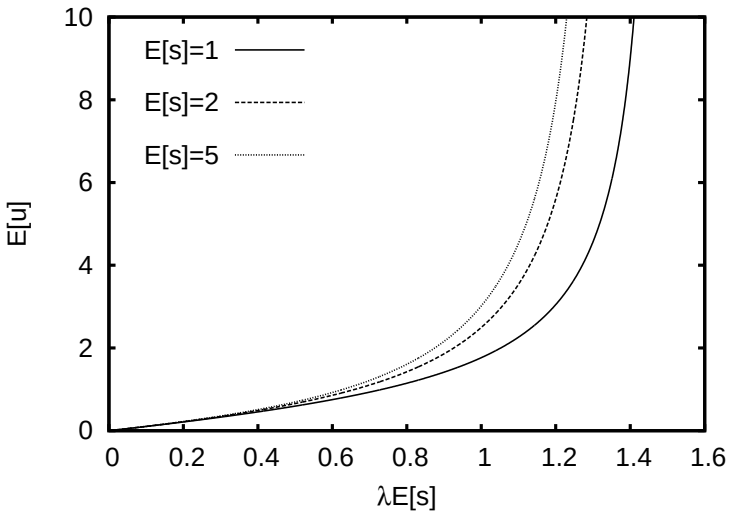

Figure 7: Mean system content $E[u]$ versus traffic intensity $\lambda E[s]$, for Poisson arrivals, $\alpha = 0.5$ and geometric service times with mean $E[s]$ as indicated.

values of $E[s]$, respectively. Both figures clearly show the expected increase of the mean system content with the traffic intensity, and exhibit vertical asymptotes at the values of $\lambda E[s]$ where the maximum achievable throughput is reached. In Fig. 6, the vertical asymptotes occur at the same values of $\lambda E[s]$, as the ones where $p(0)$ reaches the value zero in Fig. 5, i.e., the values determined from equation (86). In Fig. 7, the vertical asymptotes occur at

$$\lambda E[s] = C_{\text{geom}}(E[s]\,, 0.5) = \frac{4E[s]-1}{3E[s]-1}\ , \tag{87}$$

i.e., at traffic intensity 1.5 for $E[s] = 1$, 1.4 for $E[s] = 2$ and 1.36 for $E[s] = 5$.

5 Conclusions and future work

In this paper, we have analyzed a discrete-time queueing model with two customer classes and two class-dedicated servers, operating under the global-FCFS service discipline, assuming independent arrivals from slot to slot with a simple first-order Markovian interclass-correlation model. The paper extends earlier work ([9, 8]) from deterministic service times, either equal to 1 slot or $s > 1$ slots, to variable service times with geometric distribution. The model studied in this paper can also be considered as the discrete-time counterpart of an earlier continuous-time model ([20]) with exponential service times. The results confirm the strong impact of "class clustering" in the arrival stream on the stability and the main steady-state performance measures of a multi-class global-FCFS queueing system. In particular, they suggest a negative effect of the variability of the service times on the maximum achievable throughput of such a system. Future work could incorporate more general distributions for the service times in the model to corroborate this conjecture. Other possible extensions could consider more than two customer classes, more general interclass-correlation models and relative load distributions of the various customer classes in the aggregated arrival stream, time-correlated arrival processes, etc. Also, the derivation of the full distribution (or pgf) of customer delays and waiting times could be envisaged.

Appendix

In subsection 2.4, the last remaining unknown quantity $p(0)$ was determined by identifying a zero $z = \theta \neq 1$ of the denominator $n(z)$ in expression (37) for the steady-state pgf $U(z)$ of the system content. We now show that, in all "non-degenerate" cases, exactly one such zero always exists.

It can be verified that expression (39) for $n(z)$ can be written as

$$z^2 n(z) = \det\left(z^2\mathbf{I} - \mathbf{A}(z)\right), \tag{88}$$

with $\mathbf{I}$ the 2×2 identity matrix and

$$\mathbf{A}(z) = E(z)\begin{bmatrix}(1-\mu)^2z^2 + \mu(1-\mu)z + (1-\alpha)\mu^2 & \mu(1-\mu)z + \alpha\mu^2 \\ \mu(1-\alpha)z & (1-\mu)z^2 + \mu\alpha z\end{bmatrix}. \tag{89}$$

This is readily obtained by rewriting the set of simultaneous equations (35) and (34) as a vector-matrix equation for the row vector $\begin{bmatrix}U_0(z) & U_1(z)\end{bmatrix}$. In [13], Gail et al. studied the location and existence of zeroes in the closed unit disk of expressions like (88) in detail, and we can apply their theorems to our model. First, note that $\mathbf{A}(1)$ is the transition matrix of a 2-state discrete-time Markov chain that is irreducible. The chain is only reducible if one of the off-diagonal elements is zero, i.e., if either $\mu = 0$ or $\alpha = 1$. The first case corresponds to infinite service times while the second represents a system with one customer class only and is treated separately in section 3.2, where $p(0)$ can be obtained by direct normalization. Both cases are, in fact, degenerate instances of the system under study, and are not further considered here. Secondly, all elements of $\frac{\mathrm{d}}{\mathrm{d}z}\mathbf{A}(z)|_{z=1}$ are finite, which is equivalent to the requirement that $\lambda = E'(1)$ is finite. Thirdly, we also have that $\frac{\mathrm{d}}{\mathrm{d}z}\det(z^2\mathbf{I} - \mathbf{A}(z))|_{z=1} > 0$ since this corresponds to the stability condition (22), as can easily be verified.

Under these three conditions, it is shown in [13] that the function $z^2n(z)$ in (88) has exactly $4-g$ zeroes (counting multiplicities) inside and g zeroes on the unit circle in the complex z-plane, with $g \geqslant 1$. Obviously, $z = 0$ is a double zero, so $n(z)$ is left with $2 - g$ and g zeroes inside and on the unit circle, respectively, g being either 1 or 2. The case $g = 2$ occurs if and only if the right-hand side of (88) for $z = y^{1/2}$ is a single-valued function in y for $|y| \leqslant 1$, or in other words, if only even-degree terms z^{2k} occur in (89). Again, this only happens in the precluded cases $\mu = 0$ or $\alpha = 1$. Therefore, the conclusion is that $g = 1$ and that $n(z)$ has exactly one zero on the unit circle (this is $z = 1$) and one zero $z = \theta$ with $|\theta| < 1$.

Acknowledgements

This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (BelSPO), Belgium. Dieter Claeys is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (FWO-Vlaanderen), Belgium.

References

[1] K. Avrachenkov, E. Morozov, and B. Steyaert. Sufficient stability conditions for multi-class constant retrial rate systems. *Queueing Systems*, pages 1–23, 2015.

[2] U. Ayesta, P. Jacko, and V. Novak. Scheduling of multi-class multi-server queueing systems with abandonments. *Journal of Scheduling*, pages 1–17, 2015.

[3] S. Balsamo, G. Rossi, and A. Marin. Applying bcmp multi–class queueing networks for the performance evaluation of hierarchical and modular software systems. *International Journal of Computer Aided Engineering and Technology*, 7(2):145–157, 2015.

[4] P. Beekhuizen and J. Resing. Performance analysis of small non-uniform packet switches. *Performance Evaluation*, 66:640–659, 2009.

[5] H Bruneel. Analysis of buffer behavior for an integrated voice-data system. *Electronics Letters*, 19(2):72–74, 1983.

[6] H Bruneel. A general model for the behavior of infinite buffers with periodic service opportunities. *European Journal of Operational Research*, 16(1):98–106, 1984.

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

[8] H. Bruneel, W. Mélange, D. Claeys, S. De Vuyst, and J. Walraevens. A two-class global FCFS discrete-time queueing model with arbitrary-length constant service times. 2015.

[9] 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(1):123–132, 2012.

[10] H. Chang and G. Park. A study on traffic signal control at signalized intersections in vehicular ad hoc networks. *Ad Hoc Networks*, 11(7):2115–2124, 2013.

[11] 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.

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

[13] H. Gail, S. Hantler, and B. Taylor. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Advances in Applied Probability*, 28:114–165, 1996.

[14] J. Hsu. Buffer behavior with Poisson arrival and geometric output processes. *IEEE Transactions on Communications*, 22:1940–1941, 1974.

[15] Y. Inoue and T. Takine. The multi-class FIFO M/G/1 queue with exponential working vacations. *Journal of the Operations Research Society of Japan*, 56(2):111–136, 2013.

[16] A. Izagirre, I. Verloop, and U. Ayesta. Heavy-traffic analysis of a non-preemptive multi-class queue with relative priorities. *Probability in the Engineering and Informational Sciences*, 29(02):153–180, 2015.

[17] O. Jennings and J. Reed. An overloaded multiclass FIFO queue with abandonments. *Operations research*, 60(5):1282–1295, 2012.

[18] L. Kleinrock. *Queueing systems, part I*. Wiley, New York, USA, 1975.

[19] C. Li and M. Neely. Solving convex optimization with side constraints in a multi-class queue by adaptive c\ mu rule. *Queueing Systems*, 77(3):331–372, 2014.

[20] 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 and Management Optimization*, 10:193–206, 2014.

[21] N. Nasser, L. Karim, and T. Taleb. Dynamic multilevel priority packet scheduling scheme for wireless sensor network. *IEEE Transactions on Wireless Communications*, 12(4):1448–1459, 2013.

[22] R. Nishi, H. Miki, A. Tomoeda, and K. Nishinari. Achievement of alternative configurations of vehicles on multiple lanes. *Physical Review E*, 79:066119, 2009.

[23] D. Pandey and A. Pal. Delay analysis of a discrete-time non-preemptive priority queue with priority jumps. *Applications & Applied Mathematics*, 9(1), 2014.

[24] T. Robertazzi. *Computer networks and systems: queueing theory and performance evaluation*. Springer Science & Business Media, 2012.

[25] V. Sarhangian and B. Balcioglu. Waiting time analysis of multi-class queues with impatient customers. *Probability in the Engineering and Informational Sciences*, 27(03):333–352, 2013.

[26] D. Stanford, P. Taylor, and I. Ziedins. Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3):297–330, 2014.

[27] Ö. Ulusçu 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.

[28] A. Vadivu, R. Vinayak, S. Dharmaraja, and R. Arumuganathan. Performance analysis of voice over internet protocol via non markovian loss system with preemptive priority and server break down. *Opsearch*, 51(1):50–75, 2014.

[29] F. van Wageningen-Kessels, B. van't Hof, S. Hoogendoorn, H. Van Lint, and K. Vuik. Anisotropy in generic multi-class traffic flow models. *Transportmetrica A: Transport Science*, 9(5):451–472, 2013.

[30] T. Van Woensel and N. Vandaele. Modeling traffic flows with queueing models: A review. *Asia-Pacific Journal of Operational Research*, 24:435–461, 2007.