IMPACT OF BLOCKING WHEN CUSTOMERS OF DIFFERENT CLASSES ARE ACCOMMODATED IN ONE COMMON QUEUE

Herwig Bruneel, Willem Mélange, Bart Steyaert, Dieter Claeys and Joris Walraevens
Department of Telecommunications and Information Processing
Ghent University-UGent, Ghent, Belgium

Keywords: Queueing model, Blocking, Class clustering.

Abstract: In this paper, situations are investigated where customers requiring different types of service, each provided by distinct servers, are accommodated in one common queue. In such scenarios, customers of one class (i.e., requiring a given type of service) may be hindered ("blocked") by customers of other classes. For instance, if a road or a highway is split in two or more subroads leading to different destinations, cars on that road heading for destination A may be hindered or even blocked by cars heading for destination B, even when the subroad leading to destination A is free, simply because they have to queue in first-come-first-served (FCFS) order on the main road.

The purpose of this paper is to study the effect of blocking. We therefore develop a discrete-time queueing model and establish performance measures related to the number of waiting customers. Based on the obtained results, we demonstrate that clustering of arrivals according to class pronounces the negative impact of blocking. We believe that the impact of class clustering on blocking has been largely overlooked in the regular operations research and queueing literature.

1 INTRODUCTION

In general, when customers require some kind of service, they queue up and await their turn. This can range from people waiting at a counter of a post office to cars waiting at traffic lights. When a variety of services is provided, usually separate queues are formed for each service type. For instance, in a City Hall, different queues are created for the Register Office and the Housing Department. In some applications however, it may not be physically feasible or desirable to provide separate queues for each type of service that customers may require, and it may be necessary or desirable to accommodate different types of customers (i.e., customers requiring different types of service) in the same queue. In such cases, customers of one type (i.e., requiring a given type of service) may also be hindered by customers of other types. For instance, if a road or a highway is split in two or more subroads leading to different destinations, cars on that road heading for destination A may be hindered or even blocked by cars heading for destination B, even when the subroad leading to destination A is free, simply because they have to queue in first-come-first-served (FCFS) order on the main road. This blocking also takes place in weaving sections on highways (Ngoduy, 2006; Nishi et al., 2009). We refer to (Van Woensel and Vandaele, 2006; Van Woensel and Vandaele, 2007) for a general overview and validation of the modelling of traffic flows with queueing models. Similarly, in switching nodes of telecommunication networks, information packets with a given destination A may have to wait for the transmission of packets destined to node B that arrived earlier, even when the link to destination A is free, if the arriving packets are accommodated in so-called input queues according to the source from which they originate (the well-known HOL-blocking effect, see (Karol et al., 1987; Liew, 1994; Laevens, 1999; Stolyar, 2004; Beekhuizen and Resing, 2009)). These situations are also related to models where queues are "pooled" (see e.g. (Mandelbaum and Reiman, 1998; Van Dijk and Van der Sluis, 2008)) in the sense that customers (cars or packets) that require a different service or have a different destination share a common queue. Although the queue studied in the current paper can be considered as pooled, the difference with the models in (Mandelbaum and Reiman, 1998; Van Dijk and Van der Sluis, 2008) is that customers can be blocked by customers of the other type.

In order to gain insight into the impact of this kind of phenomenon on the performance of the involved

Bruneel H., Mélange W., Steyaert B., Claeys D. and Walraevens J..
IMPACT OF BLOCKING WHEN CUSTOMERS OF DIFFERENT CLASSES ARE ACCOMMODATED IN ONE COMMON QUEUE.
DOI: 10.5220/0003735500310038
In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems* (ICORES-2012), pages 31-38
ISBN: 978-989-8425-97-3

systems, we study the number of customers in a simple conceptual discrete-time queueing model in this paper, which is simple enough to allow explicit solution but rich enough to capture the essential aspects of the problem at hand. We envisage to analyze more general models in future work.

2 MATHEMATICAL MODEL

We consider a discrete-time queueing system with infinite waiting room, two servers, named A and B, and two types (classes) of customers, named 1 and 2. Each of the two servers is dedicated to a given class of customers, i.e., server A can only serve customers of type 1 and server B can only serve customers of type 2. Service times of all customers are deterministically equal to 1 slot each. Customers are served in their order of arrival, regardless of the class they belong to. We call this service discipline "global FCFS" in this paper.

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

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

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

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

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

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

Next, we describe the occurrence of the two types (1 and 2) in the sequence of the consecutive arriving customers. In this first study, we assume that both types of customers account for half of the total load of the system, i.e., both customer classes are equiprobable, but there may be some degree of "class clustering" in the arrival process, i.e., customers of any given type may (or may not) have a tendency to "arrive back-to-back". Mathematically, this means that the types of two consecutive customers may be non-independent. Specifically, we assume a first-order Markovian type of correlation between the types of two consecutive customers, which basically means that the probability that the next customer belongs to a given class depends on the class of the previous customer. In order to keep the model as simple as possible, 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 type as the previous one*. The parameter α can then be considered as a measure of the degree of class clustering in the arrival process, and will therefore be referred to as the "cluster parameter" in the sequel. It is easily seen that the size of a cluster of customers of the same type, i.e., the number of consecutive customers of any given type between two customers of the opposite type, is geometrically distributed with parameter α and mean value $1/(1-\alpha)$.

We note at this point that more general models could be envisaged to describe the presence of class clustering in the arrival process of the system. For instance, the transition probability to go from class 1 to 1 (e.g. α) could be chosen different from the transition probability to go from class 2 to 2 (e.g. β), and this would allow us to consider systems where the partial loads of both classes of customers are not equal, but preliminary research has revealed that this kind of generalization would complicate the analysis of the system considerably. More specifically, the analytical approach to analyze the system, as presented in the next section, would not be applicable at all. We therefore prefer to defer more general models to future work. From the conceptual point of view, the only price we pay with this choice is that we can only study cases where both classes of customers are equiprobable. The effect of class clustering on the other hand can be researched thoroughly.

It can be seen that the two-server system described above is non-workconserving, for two different (orthogonal) reasons. First, the fact that the two servers A and B are dedicated to only one type of customers each, may result in situations where only one of the servers is active even though the system contains more than one customer (of the same type, in such a case). This implies that we cannot expect the system to perform as well as a regular two-server queue with two equivalent servers, i.e., servers able to serve *all* customers. In this paper, we consider this form of inefficiency as an intrinsic feature of our system, simply caused by the fact that the customers as well as the servers are non-identical. The second reason why the system is non-workconserving lies in the use of the global FCFS service discipline. This rule may result in situations where only one server is active although the system contains customers of *both* classes. Such situations occur whenever the two "eldest" customers in the system, i.e., the two customers at the front of the queue, are of the same type: only one of them can then be served (by its own dedicated server) and the other "blocks" the access to the second server for

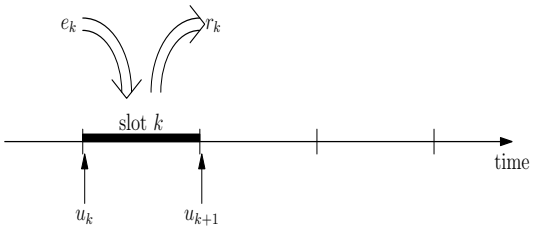

Figure 1: Time axis to illustrate the system equations.

customers of the opposite type further in the queue. This second form of inefficiency is not an intrinsic feature of two-class systems with dedicated servers, but rather it is due to the accidental order in which customers of both types happen to arrive (and receive service) in the system (as described by the parameter α in our model). It is this second mechanism that we want to emphasize in the paper.

The structure of the rest of this paper is as follows. Section 3 first presents a general analysis of the number of customers in the system: an expression is derived for the pgf of this number and a method is described to determine the two remaining unknowns in that expression. Next, for the special case of geometric arrivals, explicit closed-form expressions are obtained not only for the pgf but also for the pmf and the mean value of the number of customers in the system. We discuss the results both conceptually and quantitatively in section 4. Some conclusions are drawn and directions for future work are given in section 5.

3 SYSTEM ANALYSIS

3.1 System Equations

We start the analysis by defining a number of important random variables, illustrated in Fig. 1. Specifically, let u_k denote the total system occupancy, i.e., the total number of customers present in the system at the beginning of the k-th slot, and e_k the total number of arrivals in the system during this slot (with known pmf $e(n)$ and pgf $E(z)$). Furthermore, let r_k (initially) denote the number of customers served during the k-th slot, when $u_k > 1$. Then the following recursive system equations can be established:

$$u_{k+1} = e_k \quad , \quad \text{if } u_k \leq 1 \ ,$$
$$u_{k+1} = u_k + e_k - r_k \quad , \quad \text{if } u_k > 1 \ .$$

The two above cases can be summarized in one single system equation

$$u_{k+1} = e_k + (u_k - r_k)^+ \quad , \tag{1}$$

by introducing the notation $(\ldots)^+$ to indicate the quantity $\max(0,\ldots)$. In equation (1), the random variables $\{r_k\}$ can be treated as a sequence of strictly positive i.i.d. random variables (indicating the numbers of "available servers" during the consecutive slots) with common pmf

$$r(n) \triangleq \text{Prob}[r_k = n] \quad , \quad 1 \leq n \leq 2 \ ,$$

and common pgf

$$R(z) \triangleq \sum_{n=1}^{2} r(n) z^n \ ,$$

whereby

$$r(1) = \alpha \ , \quad r(2) = 1 - \alpha \ ,$$

and

$$R(z) = \alpha z + (1-\alpha) z^2 \ . \tag{2}$$

In fact, this observation is the key to the solution. It actually turns out that the number of customers that can be served in slot k (with $u_k > 1$) does not depend on the actual type of the customer in the head-of-line position, but only on the identity or non-identity of the classes to which the two "eldest" customers (at the front of the queue) belong, regardless of the numbers of customers served during previous slots. If both customers belong to the same class, which happens with probability α, irrespective of the type of the head-of-line customer, then only one customer can be served. If the two customers belong to opposite classes, then both will be served; this case occurs with probability $1-\alpha$. This explains why $r(1) = \alpha$ and $r(2) = 1-\alpha$. It is clear that equation (1) is also correct if $u_k \leq 1$, because, with the given definition of the r_k's, $(u_k - r_k)^+$ is equal to zero in such cases.

The fact that the random variables $\{r_k\}$ are independent, in spite of the correlated nature of the types of consecutive customers, stems from the fact that - in this particular model - the probability that the next customer has the same type as the previous customer is simply given by the cluster parameter α, regardless of the specific type of the previous customer. This simplifying circumstance does not exist in more general models for the "class clustering" mechanism.

3.2 Analysis of the System Occupancy

For all k, let $U_k(z)$ denote the pgf of u_k. Then, from equation (1) we can derive

$$U_{k+1}(z) = E(z) \cdot E\left[z^{(u_k - r_k)^+}\right] \quad , \tag{3}$$

with $E[\cdot]$ the expectation operator. The second factor in the right hand side of (3) can be expanded further

by means of the law of total probability (using also the mutual independence of u_k and r_k):

$$E\left[z^{(u_k-r_k)^+}\right] = \alpha E\left[z^{(u_k-1)^+}\right] + (1-\alpha)E\left[z^{(u_k-2)^+}\right] \ . \quad (4)$$

Here, the two remaining expectations are to be taken with respect to one single random variable u_k. Using standard z-transform techniques in equation (4), and combining the result with (3), we then obtain

$$U_{k+1}(z) = E(z)\cdot\Bigg(R(1/z)U_k(z) + \frac{z-1}{z^2}\big[(z+1-\alpha)u_k(0) + (1-\alpha)zu_k(1)\big]\Bigg) \ , \quad (5)$$

where, for all $i \geq 0$,

$$u_k(i) \triangleq \text{Prob}[u_k = i] \ .$$

Now, let us assume that the queueing system at hand is stable, i.e., that the stability condition is fulfilled. It is not difficult to see that, with the system equations established above, the system is stable if and only if the mean number of arrivals per slot, given by $E'(1)$, is strictly less than the mean number of available servers per slot, given by $R'(1)$, i.e., if and only if

$$E'(1) < R'(1) \ ,$$

or, expressed in the basic parameters of our system,

$$\lambda < 2-\alpha \ . \quad (6)$$

We now let the time parameter k go to infinity. Assuming the system reaches a steady state, then both functions $U_k(z)$ and $U_{k+1}(z)$ converge to a common limit function $U(z)$, which denotes the pgf of the system occupancy at the beginning of an arbitrary slot in steady state. As a result, equation (5) translates into a linear equation for $U(z)$, with solution

$$U(z) = \frac{(z-1)E(z)\,[u(0)(z+1-\alpha)+u(1)(1-\alpha)z]}{z^2-(1-\alpha+\alpha z)E(z)} \ , \quad (7)$$

where

$$u(i) \triangleq \lim_{k\to\infty} u_k(i) \ .$$

This expression contains only known quantities, except for the two unknown probabilities $u(0)$ and $u(1)$. These can be determined, in general, by invoking the well-known property that pgf's such as $U(z)$ are bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, at least when the stability condition (6) of the queueing system is met (only in such a case our analysis was justified and $U(z)$ can be viewed as a legitimate pgf). Now, it can be shown by means of Rouché's theorem from complex analysis (Gonzáles, 1992; Bruneel and Kim, 1993) that the denominator of equation (7) has exactly two zeroes inside the closed unit disk of the complex z-plane, one of which is equal to 1, as soon as the stability condition (6) is fulfilled. It is clear that these two zeroes should also be zeroes of the numerator of equation (7), as $U(z)$ must remain bounded in those points. For the zero $z = 1$, this condition is fulfilled regardless of the values of the unknowns $u(0)$ and $u(1)$, since the numerator of (7) contains a factor $z-1$. However, for the second zero, the requirement that the numerator should vanish yields a linear equation for the two unknowns. A second linear equation can be obtained by invoking the normalizing condition of the pgf $U(z)$, i.e., the condition $U(1) = 1$. In general, the two unknown probabilities $u(0)$ and $u(1)$ can be found as the solutions of the two established linear equations. Substitution of the obtained values in equation (7) then leads to a fully determined expression of the steady-state pgf $U(z)$ of the system occupancy.

From this result, various performance measures of practical importance can then be derived. For instance, the mean system occupancy can be found as $E[u] = U'(1)$. By applying (the discrete-time version of) Little's result (Kleinrock, 1975; Bruneel and Kim, 1993; Fiems and Bruneel, 2002), the mean delay (system time) of a customer can be obtained as $E[d] = U'(1)/\lambda$, and so on. In the next subsection, we treat a special case in which the computations can be further simplified and explicit closed-form expressions can be obtained for most quantities of interest.

3.3 Special Case: Geometric Arrivals

Let us consider the special case whereby the number of arrivals per slot has a geometric distribution with mean value λ. Then, $e(n)$ and $E(z)$ are given by

$$e(n) = \frac{1}{1+\lambda}\left(\frac{\lambda}{1+\lambda}\right)^n \ , \quad n \geq 0 \ ,$$

$$E(z) = \frac{1}{1+\lambda-\lambda z} \ ,$$

and (7) can be rewritten as

$$U(z) = \frac{u(0)(z+1-\alpha)+u(1)(1-\alpha)z}{-\lambda z^2+z+(1-\alpha)} \ , \quad (8)$$

where we have cancelled out a common factor $z-1$ from the numerator and the denominator.

It is not difficult to see that, as soon as the stability condition (6) is satisfied, the (quadratic) denominator of (8) has two zeroes, one of which (z_1) is inside the

unit disk, and one of which (z_0) is outside the unit disk. As explained above, the bounded nature of $U(z)$ inside the unit disk implies that z_1 should also be a zero of the numerator of equation (8), which happens to be a linear function of z. It then follows that $U(z)$ can be further simplified by cancelling out the common factor $z - z_1$ from the numerator and the denominator and using the normalizing condition $U(1) = 1$. As a result we obtain

$$U(z) = \frac{1 - z_0}{z - z_0} \ , \qquad (9)$$

where z_0 is given by

$$z_0 = \frac{1 + \sqrt{1 + 4\lambda(1-\alpha)}}{2\lambda} \ . \qquad (10)$$

The pgf $U(z)$ given in equation (9) can be easily inverted; the corresponding pmf of the steady-state system occupancy reads

$$u(i) = \left(1 - \frac{1}{z_0}\right)\left(\frac{1}{z_0}\right)^i \ , \quad i \geq 0 \ , \qquad (11)$$

i.e., the system occupancy has a geometric distribution with parameter $1/z_0$. The tail distribution $\text{Prob}[u > i]$, i.e., the probability that more than i customers be present in the system — which can be used as a rough approximation for the loss probability in a finite-capacity system with room for exactly i customers, see (Steyaert and Bruneel, 1995; Gouweleeuw and Tijms, 1998; Kim and Schroff, 2001) — can be expressed as

$$\text{Prob}[u > i] = \left(\frac{1}{z_0}\right)^{i+1} \ , \quad i \geq 0 \ . \qquad (12)$$

The mean system occupancy $E[u]$ at the beginning of an arbitrary slot can be easily derived as well:

$$E[u] = \frac{1}{z_0 - 1} = \frac{1 - 2\lambda - \sqrt{1 + 4\lambda(1-\alpha)}}{2(\lambda - 2 + \alpha)} \ . \qquad (13)$$

Finally, the mean delay $E[d]$ of a customer (expressed in time slots) can be obtained from the discrete-time version of Little's result (Bruneel and Kim, 1993; Fiems and Bruneel, 2002):

$$E[d] = \frac{E[u]}{\lambda} = \frac{1 - 2\lambda - \sqrt{1 + 4\lambda(1-\alpha)}}{2\lambda(\lambda - 2 + \alpha)} \ . \qquad (14)$$

It is worth noting that the stability condition (6) is clearly reflected in the expressions (13) and (14), in that the denominators of both expressions tend to zero as the mean arrival rate λ approaches its limiting value $2 - \alpha$, indicating the unbounded growth of (mean) buffer occupancy and delay as the system approaches the border of its stability region.

4 DISCUSSION OF RESULTS AND NUMERICAL EXAMPLES

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

The first interesting result obtained is the form of the stability condition (6),

$$\lambda < 2 - \alpha \ ,$$

which shows that the maximum achievable throughput of this system, expressed in customers per slot, is very directly determined by the degree of class clustering in the arrival process as described by the cluster parameter α. For this specific model, the formula is remarkably simple and shows that the achievable throughput decreases linearly with the cluster parameter α. As α can take values between 0 and 1, the maximum throughput can vary between (nearly) 2 customers per slot and (nearly) 1 customer per slot. It is interesting to look at the extreme values $\alpha = 0$ and $\alpha = 1$. If the cluster parameter is equal to zero, then the types of two consecutive customers are always opposite, and one type of customers can never block the other type; in this case both servers A and B are active as soon as at least two customers are present in the system, i.e., the system is work-conserving and behaves as a regular queue with two identical servers able to serve all customers. However, as soon as some amount of "class clustering" appears in the arrival stream, the achievable throughput is affected, according to equation (6). In the extreme case where the cluster parameter is equal to 1, all customers belong to the same class and only one of the two servers is actually being used by the arrival stream; in this case, the system behaves as a single-server queue and the throughput can never exceed 1 customer per slot.

These results show that the presence of "class clustering" in the arrival stream of a multiclass queue with dedicated servers and "global FCFS" service discipline can actually be devastating for the performance of the queue, and we believe that this phenomenon has been largely overlooked in the regular queueing literature. Another way of looking at this phenomenon is to rewrite the inequality (6) as

$$\lambda + \alpha < 2 \ , \qquad (15)$$

which seems to say that the actual traffic intensity (λ) and the cluster parameter (α) are equally important with respect to the stability of the queue: you can afford more load only if you can decrease the class clustering of the arrival stream, i.e., the class clustering appears to represent some kind of additional or virtual load to the system. In this sense, the quantity

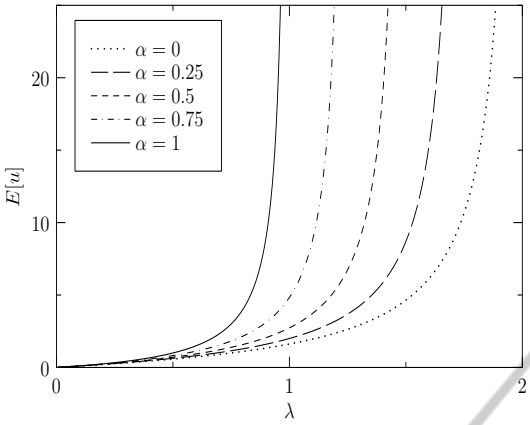

Figure 2: Mean system contents versus the mean arrival rate for various values of the cluster parameter.

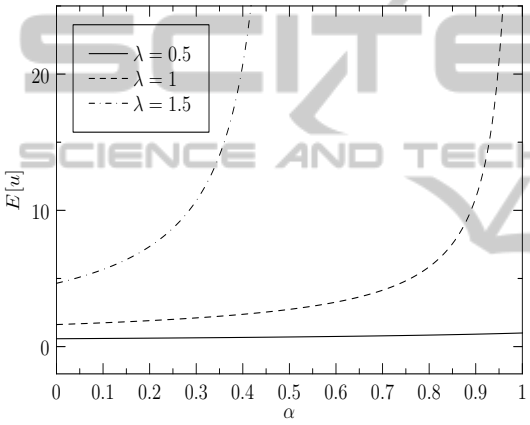

Figure 3: Mean system contents versus the cluster parameter for various values of the mean arrival rate.

$\lambda + \alpha$ could be considered as some kind of equivalent traffic intensity of the system.

For the case of geometric arrivals, as discussed in subsection 3.3, we show some numerical results in figures $2-4$.

Fig. 2 shows the mean system contents $E[u]$ versus the mean arrival rate λ, for various values of the cluster parameter α. The figure clearly illustrates the great and direct (negative) impact of "class clustering" on the average number of customers in the system, for any given arrival intensity lower than 1. More generally, it also shows the shrinking stability region of the system, as the degree of class clustering increases. We note that the value $\alpha = 0.5$ represents the case where the types of consecutive customers in the arrival stream are independent. Our results prove that neglecting the correlation between the types of consecutive customers may lead to either serious underestimation or overestimation of the mean system occupancy.

In Fig. 3, we have plotted the mean system contents $E[u]$ versus the cluster parameter α, for given values of the arrival rate λ. The figure shows that for lightly loaded systems (e.g. $\lambda = 0.5$ in the figure) the influence of class clustering is negligible. This is also intuitively clear: the demand of the arrival stream, in such a case, is considerably less than the traffic that can be handled by 1 server, and therefore, the question of whether the second server is also active or not — which is determined by the amount of class clustering — is not very relevant. However, as soon as the arrival rate λ exceeds the value 1, the cluster parameter α becomes important. Specifically, the average queue size can even grow without bound when α reaches the value $2-\lambda$.

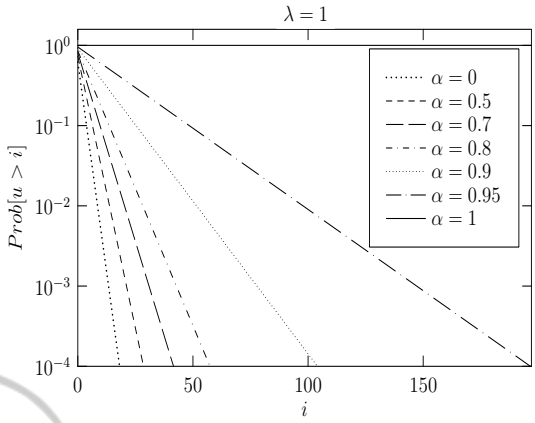

Figure 4: Tail probability of the system contents for a given arrival rate of 1 and various values of the cluster parameter.

Fig. 4 shows the tail probability $\text{Prob}[u > i]$, which can be considered as an approximate value for the loss probability in a system with finite storage capacity equal to i places, versus the value of i, for a given value $\lambda = 1$ and various values of the cluster parameter α. The results in this figure can be used, for instance, for dimensioning purposes of the required buffer size to achieve a prescribed loss ratio. As an example, let us assume a target loss ratio of 10^{-4}, then the graphs in Fig. 4 show that the required buffer size depends very strongly on the cluster parameter α: for $\alpha = 0$, a storage capacity of 18 is sufficient; $\alpha = 0.5, 0.7, 0.8, 0.9$ and 0.95 require a buffer size of 29, 42, 58, 105, 197 respectively, whereas for $\alpha = 1$ the system is unstable and a loss ratio of 10^{-4} is not even achievable.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have developed a queueing model that enables to evaluate scenarios where customers

requiring different service types, each provided by distinct servers, are accommodated in one common queue. We have proposed a dual-class, two-server queueing model with class-dedicated servers in discrete time, operating under the global FCFS service discipline, assuming independent arrivals from slot to slot with a simple first-order Markovian class clustering model. The model is relatively simple so as to allow for an analytical solution, but yet contains all the important elements needed for a conceptual study of the effect of "global FCFS" on this type of queue. We emphasize that we have succeeded in deriving an explicit formula for the pgf of the system occupancy, under general assumptions with respect to the arrival statistics. For the special case of geometric arrivals, we have even been able to obtain explicit closed-form expressions for the pmf's, the mean values and the tail distributions of system occupancy. The results reveal the very direct and great influence of the degree of "class clustering" in the arrival stream on the stability and the performance of the system. We believe that this is the main qualitative conclusion of the study.

In general, only few studies have focused on the phenomenon of class clustering in the context of multiclass queueing systems, and this paper shows that the effect of class clustering may be very important, possibly not only in queues with class-dependent servers and global FCFS, but also in other queueing situations whereby the service mechanism is sensitive to the order of service of customers of different classes. For instance, we expect that class clustering may also have substantial effects on the performance of priority queues or queues whereby the lengths of the service times depend on the way customers of different types succeed each other.

The model examined in this paper can be generalized in various directions. To start with, more general service-time distributions can be considered than the simple deterministic one-slot-per-customer model studied in this paper. The extra difficulty is that two customers of different types do not necessarily leave the system in order of their arrival anymore, exactly due to the variable service times. We note that the simplest model in continuous time with exponential service times already has this difficulty (see e.g. (Mélange et al., 2011)). Also, the assumption of independent arrivals from slot to slot may be relaxed to allow for correlated or bursty types of arrival processes. Depending on the precise details of the class clustering model used, this may also affect the performance of the system considerably. Finally, we may consider more complicated models for the class clustering mechanism than the simple one-parameter model used in this paper. As already touched upon in section 2, the current study is restricted to systems whereby both customer classes are equiprobable and the probability of having a next customer of the same (or opposite) type as the previous one does not depend on the type of the previous customer. Many different types of more general models than this can be envisaged. For instance, the types of consecutive customers in the arrival stream could be modeled as a two-state Markov chain with general transition probabilities, or the sizes of subsequent clusters of customers of each type could be described by general (rather than geometric) probability distributions, and so on. We plan to tackle several of these generalizations in future work.

REFERENCES

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

Bruneel, H. and Kim, B. (1993). *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA.

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

Gonzáles, M. (1992). *Classical complex analysis*. Marcel Dekker, New York, USA.

Gouweleeuw, F. and Tijms, H. (1998). Computing loss probabilities in discrete-time queues. *Operations Research*, 46:149–154.

Karol, M., Hluchyj, M., and Morgan, S. (1987). Input versus output queueing on a space-division packet switch. *IEEE Transactions on Communications*, 35:1347–1356.

Kim, H. and Schroff, N. (2001). Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. on Networking*, 9:755–768.

Kleinrock, L. (1975). *Queueing systems, part I*. Wiley, New York, USA.

Laevens, K. (1999). A processor-sharing model for input-buffered ATM-switches in a correlated traffic environment. *Microprocessors and Microsystems*, 22:589–596.

Liew, S. (1994). Performance of various input-buffered and output-buffered ATM switch design principles under bursty traffic: simulation study. *IEEE Transactions on Communications*, 42:1371–1379.

Mandelbaum, A. and Reiman, M. (1998). On pooling in queueing networks. *Management Science*, 44:971–981.

Mélange, W., Bruneel, H., Steyaert, B., and Walraevens, J. (2011). A two-class continuous-time queueing model with dedicated servers and global fcfs service discipline. In *Analytical and Stochastic Modeling Techniques and Applications*, volume 6751 of *Lecture*

Notes in Computer Science, pages 14–27. Springer Berlin / Heidelberg.

Ngoduy, D. (2006). Derivation of continuum traffic model for weaving sections on freeways. *Transportmetrica*, 2:199–222.

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

Steyaert, B. and Bruneel, H. (1995). *Accurate approximation of the cell loss ratio in ATM buffers with multiple servers*, volume 1 of *Performance Modelling and Evaluation of ATM Networks*, pages 285–296. Chapman & Hall, London.

Stolyar, A. (2004). MaxWeight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, 14:1–53.

Van Dijk, N. and Van der Sluis, E. (2008). To pool or not to pool in call centers. *Production and Operations Management*, 17:1–10.

Van Woensel, T. and Vandaele, N. (2006). Empirical validation of a queueing approach to uninterrupted traffic flows. *4OR, A Quarterly Journal of Operations Research*, 4:59–72.

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