Effect of global FCFS and relative load distribution in two-class queues with dedicated servers

Herwig Bruneel Willem Mélange Bart Steyaert Dieter Claeys
Joris Walraevens

Department of Telecommunications and Information Processing
Ghent University - UGent
E-mail: {hb,wmelange,bs,dc,jw}@telin.UGent.be

Abstract

In this paper, we investigate multi-class multi-server queueing systems with global FCFS policy, i.e., where customers requiring different types of service - provided by distinct servers - are accommodated in one common FCFS queue. In such scenarios, customers of one class (i.e., requiring a given type of service) may be hindered by customers of other classes. The purpose of this paper is twofold: to gain (qualitative and quantitative) insight into the impact of (i) the global FCFS policy and (ii) the relative distribution of the load amongst the customer classes, on the system performance. We therefore develop and analyze an appropriate discrete-time queueing model with general independent arrivals, two (independent) customer classes and two class-specific servers. We study the stability of the system and derive the system-content distribution at random slot boundaries; we also obtain mean values of the system content and the customer delay, both globally and for each class individually. We then extensively compare these results with those obtained for an analogous system without global FCFS policy (i.e., with individual queues for the two servers). We demonstrate that global FCFS, as well as the relative distribution of the load over the two customer classes, may have a major impact on the system performance.

Key words: multi-class queues; dedicated servers; global FCFS; relative load distribution

1 Introduction

We study a discrete-time queueing system with infinite waiting room, two types (classes) of customers, each class having its own dedicated server. Service times of all customers are deterministically equal to 1 slot each. All customers are accommodated in a common queue and are served in their order of arrival, regardless of the class they belong to, i.e., the service discipline is "global FCFS". Dedicated service and a common queue can arise in various contexts, ranging from weaving sections on highways [15, 16] to input queueing in packet switches [2, 13, 19, 24]. We refer to [6] for more details of those applications.

The purpose of this paper is twofold: to gain (qualitative and quantitative) insight into the impact of (i) the global FCFS policy and (ii) the relative distribution of the load among the customer classes, on the system performance. We study the stability of the system and the system-content distribution at random slot boundaries; we also obtain mean values of the system content and the customer delay, both globally and for each class individually. We then compare these results with those obtained for an analogous system without global FCFS policy (i.e., with individual queues for the two servers).

The model under consideration is fundamentally different from traditional multi-class systems. In traditional multi-class systems, customers of different classes compete for the *same resources* [1, 8, 10, 12, 17, 22, 23]. In the current system on the other hand, each server is dedicated to one specific class of customers. A major consequence is that traditional multi-class systems are work conserving, whereas the system described here is non-work-conserving, 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-work-conserving 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 are of the same type: only one of them can then be served (by its own dedicated server), and the other is at the front of the queue and "blocks" the access to the second server for 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. It is this second mechanism that we want to emphasize in the paper.

Although global FCFS is related to resource sharing, there are fundamental differences. In traditional resource sharing, systems with the *same* type of customers and servers are merged. In this paper, resource sharing refers to sharing a buffer instead of servers, and our servers are dedicated to specific classes of customers. In addition, traditional resource sharing is motivated from efficiency reasons (see e.g. [20]), whereas a common queue arises when it is not physically feasible or desirable to provide separate queues and can have a detrimental effect on the system performance.

There exists some overlap between the current paper[1] and our conference paper [4] and its extended version [6], where we also consider global FCFS. In [4] and [6] however, we focus on the influence of "class clustering" (i.e., the tendency of customers of equal classes to arrive back-to-back) under the assumption of symmetric load of the customer classes. In the current paper, we omit the assumption of symmetric load, and we investigate the impact of the relative load distribution on the system performance. This paper is also related with our conference paper [14], where we also focus on the combination of global FCFS and relative load distribution. As compared to [14], we here consider a discrete-time queueing model, which is more natural in a telecommunications context, where operations are synchronized to the system clock. In addition, discrete-time systems can be used to approximate continuous-time systems, but the reverse is generally not true [21]. A second major difference as compared to [14], is that we here consider deterministic instead of exponential service times. In continuous time, studying deterministic service times is a lot more difficult as compared to exponential service times, due to the loss of the memoryless property of the exponential distribution. Considering a discrete-time queueing model not only is more natural in a clocked context, but also allows a more tractable analysis in case of deterministic service times, by letting the slot length correspond with the length of a service time.

The remainder of the paper is organized as follows: in section 2 the investigated model is described in detail. The stability of the system and the system-content and delay characteristics at random slot boundaries are analyzed in section 3. Analogous results in case of individual queues (no global FCFS) are briefly summarized in section 4. Section 5 discusses some numerical results, in order to quantitatively determine the impact of the global FCFS policy and the relative distribution of the load on the performance of the system. Finally, some conclusions are drawn and directions for future research are given in section 6.

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 A and B. Each of the two servers is dedicated to a given class of customers, i.e., server A can only serve customers of type A and server B can only serve customers of type B. 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, i.e., the service discipline is "global FCFS".

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 independent and identically distributed (i.i.d.) discrete random variables with common probability mass function (pmf) $e(n)$ and common probability generating function (pgf) $E(z)$ respectively. The (total) mean number of arrivals per slot, in the sequel referred to as the (total) mean arrival rate, is given by

$$\lambda \triangleq E'(1) \ .$$

Next, we describe the relative distribution of the arrival stream (and thus of the load) amongst the customer classes. We assume that an arriving customer belongs to class A with

[1] This paper is an extended version of our conference paper [5].

probability σ, and to class B with probability $1-\sigma$, independently from customer to customer. The mean per-class arrival rates, λ_A for class A and λ_B for class B, are then given by

$$\lambda_A = \sigma\lambda \quad ; \quad \lambda_B = (1-\sigma)\lambda \ .$$

3 System analysis

3.1 Stability

Let us observe the behaviour of the system at a random slot in a nearly-saturated condition. It is clear that, regardless of the types of the customers in the queue, at least one (i.e., the eldest) customer leaves the system at the end of the slot. A second (i.e., the second eldest) customer might also leave the system, but this is dependent on the equality of its type with the type of the customer that certainly leaves the system. As a result, the mean number of customers that can leave the system at the end of the slot can be expressed as

$$1 + \text{Prob[Two eldest customers are of opposite types]} \ .$$

It thus remains to determine the probability that the two eldest customers are of opposite types. In that regard, one should be very careful: it would be premature to think that this probability equals the sum of the probability that a random customer is of type A and the next of type B and the probability that a random customer is of type B and the next of type A, leading to $2\sigma(1-\sigma)$. This is incorrect! The reason is that the types of the two eldest customers during consecutive slots are non-independent. If only one customer can be served in a slot and this customer is of type A (thus served by server A), this entails that the second eldest customer is also of type A (otherwise server B would be processing a customer of type B). As a result, the eldest customer in the next slot is of type A. A similar reasoning holds when the server can only serve one customer in a slot and this customer is of type B. Note, on the other hand, that if both servers are available, this produces no information about the type of the eldest customer in the next slot.
Taking into account the above observations, the active servers during consecutive slots can be characterized by a Markov chain with state space

- $(1, A)$: server A is active, server B is idle
- $(1, B)$: server B is active, server A is idle
- (2): both servers are active

and with state transition diagram as depicted in Fig. 1.
Note that, as explained, no direct transitions are possible between the states $(1, A)$ and $(1, B)$. The Markov chain has the following equilibrium distribution:

$$p_{1,A} = \frac{\sigma^3}{1-2\sigma(1-\sigma)} \ ,$$

$$p_{1,B} = \frac{(1-\sigma)^3}{1-2\sigma(1-\sigma)} \ ,$$

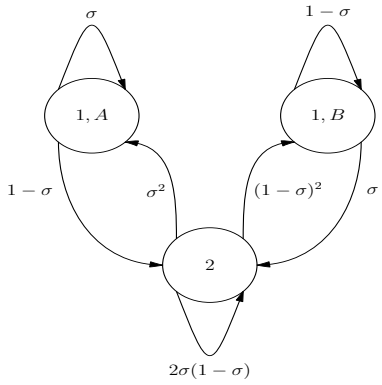

Figure 1: State transition diagram of the Markov chain of the servers' activity in a nearly saturated system

$$p_2 = \frac{\sigma(1-\sigma)}{1-2\sigma(1-\sigma)} \ ,$$

where $p_{1,A}$, $p_{1,B}$ and p_2 represent the equilibrium probabilities of finding the servers in state $(1, A)$, $(1, B)$ or (2) respectively, during a random slot. The mean number of customers that can be processed during a random slot is thus given by

$$1 + p_2 = 1 + \frac{\sigma(1-\sigma)}{1-2\sigma(1-\sigma)} \ .$$

As a result, the stability condition of the system reads

$$\lambda < 1 + \frac{\sigma(1-\sigma)}{1-2\sigma(1-\sigma)} \ . \tag{1}$$

3.2 System state description

Let us now consider a non-saturated system, assuming that (1) is satisfied. Whenever the system was empty or contained exactly one customer (two cases) at the previous slot mark, then at the current slot boundary the system contains only those customers that have arrived during the previous slot. As the type of a customer is independent of the types of previous customers, the type of the customer served (if any) during the previous slot has no influence on the types of the customers present at the current slot mark and consequently has no influence on the number of customers that can be served during the current slot. This implies that, in these two cases, the system state does not require any class-related information. If, on the other hand, multiple customers were present at the previous slot mark, then the types of the two eldest customers at the previous slot boundary may have an influence on the type of the eldest customer at the current slot boundary. Indeed, when the two eldest customers in the previous slot were of the same type, the eldest customer in the current slot was the second eldest customer in the previous slot and is thus of the same type. This, in turn, implies that the system state needs to include (at least) information on the type of the eldest customer, in these cases. In view of these observations, the evolution of the system from slot to slot can be described by a Markov chain with state space

$$\{(0), (1), (i, n), i \in \{A, B\}, n \geq 2\} \ ,$$

where (0) represents an empty system, (1) denotes a system containing one customer and (i, n) characterizes a system with n customers, the eldest customer being of type i.

3.3 System-content analysis

Let u_k represent the system content (number of customers in the system, including the customers in service, if any) at slot mark k. The pgf of the system content at a random slot boundary in steady state is denoted by $U(z)$. Next, we indicate the type of the eldest customer in slot k, when at least two customers are present, by t_k. The steady-state probabilities corresponding to the states (0), (1), (i, n) are designated by p_0, p_1 and $p(i, n)$, respectively, i.e.,

$$p_n \triangleq \lim_{k\to\infty} \text{Prob}[u_k = n] \quad , \quad n \in \{0, 1\} \ ,$$

$$p(i, n) \triangleq \lim_{k\to\infty} \text{Prob}[t_k = i, u_k = n] \quad , \quad i \in \{A, B\}, n \geq 2 \ .$$

We then have that

$$U(z) = p_0 + p_1 z + Q_A(z) + Q_B(z) \ ,$$

where

$$Q_i(z) \triangleq \sum_{n=2}^{\infty} p(i, n) z^n \quad , \quad i \in \{A, B\} \ .$$

We now calculate $Q_A(z)$ and $Q_B(z)$. Therefore, we start from the balance equation for state (A, n), $n \geq 2$:

$$\begin{aligned} p(A, n) = \ & p_0 \sigma e(n) + p_1 \sigma e(n) \\ & + \sum_{m=2}^{n+2} \left[\sigma e(n - m + 1) + (1 - \sigma)\sigma e(n - m + 2)\right] p(A, m) \\ & + \sum_{m=2}^{n+2} \sigma^2 e(n - m + 2) p(B, m) \ , \end{aligned} \tag{2}$$

where $e(-1) \triangleq 0$. Multiplying both sides of (2) by z^n, then summing over n from 2 to infinity and taking into account the definition of $Q_A(z)$ yields

$$\begin{aligned} Q_A(z) = \ & \sigma(p_0 + p_1)\left[E(z) - e(0) - e(1)z\right] \\ & + \sigma \sum_{m=2}^{\infty} p(A, m) \sum_{n=\max(2,m-1)}^{\infty} e(n - m + 1) z^n \\ & + (1 - \sigma)\sigma \sum_{m=2}^{\infty} p(A, m) \sum_{n=\max(2,m-2)}^{\infty} e(n - m + 2) z^n \end{aligned}$$

$$+\sigma^2\sum_{m=2}^{\infty}p(B,m)\sum_{n=\max(2,m-2)}^{\infty}e(n-m+2)z^n \ .$$

Relying on the definitions of $Q_A(z)$ and $Q_B(z)$ produces

$$\begin{aligned}\left[z^2-\sigma(1-\sigma)E(z)\ -\sigma zE(z)\right]Q_A(z)-\sigma^2E(z)Q_B(z)&\\ =\ \sigma(p_0+p_1)z^2\left[E(z)-e(0)-e(1)z\right]-\sigma p(A,2)e(0)z^3&\\ -\sigma z^2[(1-\sigma)p(A,2)+\sigma p(B,2)][e(0)+e(1)z]&\\ -\sigma z^3[(1-\sigma)p(A,3)+\sigma p(B,3)]e(0)\ .&\end{aligned}\tag{3}$$

Before proceeding, we note that the balance equation for state 1 reads

$$\begin{aligned}p_1=\ p_0e(1)+p_1e(1)+p(A,2)\sigma e(0)+p(B,2)\sigma e(1)+p(A,2)(1-\sigma)e(1)&\\ +p(B,2)(1-\sigma)e(0)+p(A,3)(1-\sigma)e(0)+p(B,3)\sigma e(0)\ ,&\end{aligned}$$

or equivalently

$$\begin{aligned}[(1-\sigma)p(A,3)+\sigma p(B,3)]e(0)=-\ p_0e(1)+p_1(1-e(1))-p(A,2)[\sigma e(0)+(1-\sigma)e(1)]&\\ -\ p(B,2)[\sigma e(1)+(1-\sigma)e(0)]\ .&\end{aligned}$$

Invoking this relationship in (3) yields the following linear relation between $Q_A(z)$ and $Q_B(z)$:

$$\begin{aligned}\left[z^2-\sigma(1-\sigma)E(z)\ -\sigma zE(z)\right]Q_A(z)-\sigma^2E(z)Q_B(z)&\\ =\ \sigma p_0z^2\left[E(z)-e(0)\right]+\sigma p_1z^2\left[E(z)-e(0)-z\right]&\\ -\sigma p(B,2)e(0)z^2[\sigma-(1-\sigma)z]-\sigma(1-\sigma)p(A,2)e(0)z^2(z+1)\ .&\end{aligned}\tag{4}$$

Along the same lines as above, a second equation for $Q_A(z)$ and $Q_B(z)$ can be deduced:

$$\begin{aligned}\left[z^2-\sigma(1-\sigma)E(z)-(1-\sigma)zE(z)\right]Q_B(z)-(1-\sigma)^2E(z)Q_A(z)&\\ =\ \ (1-\sigma)p_0z^2\left[E(z)-e(0)\right]+(1-\sigma)p_1z^2\left[E(z)-e(0)-z\right]&\\ -(1-\sigma)p(A,2)e(0)z^2[(1-\sigma)-\sigma z]-\sigma(1-\sigma)p(B,2)e(0)z^2(z+1)\ .&\end{aligned}\tag{5}$$

The unknown partial generating functions $Q_A(z)$ and $Q_B(z)$ can now be found by solving the set of linear (algebraic) equations (4) and (5), which yields

$$\begin{aligned}Q_A(z)=\sigma z^2\Big[\Big\{-(1-e(0))z^2+[\{1+(1-\sigma)(1-e(0))\}E(z)-1]z&\\ +E(z)[1-\sigma e(0)-(1-\sigma)E(z)]\Big\}p_0&\end{aligned}$$

$$
\begin{aligned}
&+\Big\{-(1-e(0))z^2+\{1+(1-\sigma)(1-e(0))\}zE(z)\\
&-E(z)\left[\sigma e(0)+(1-\sigma)E(z)\right]\Big\}p_1\\
&+e(0)\Big\{z^2-\{(1-\sigma)z+\sigma\}E(z)\Big\}p(B,2)\Big]\\
&/\left[z^3-z^2E(z)+\sigma(1-\sigma)zE(z)(E(z)-2)+\sigma(1-\sigma)E(z)^2\right] \quad , \qquad (6)
\end{aligned}
$$

$$
\begin{aligned}
Q_B(z)=(1-\sigma)z^2\Big[&\Big\{-(1-e(0))z^2+[\{1+\sigma(1-e(0))\}E(z)-1]z\\
&+E(z)[1-(1-\sigma)e(0)-\sigma E(z)]\Big\}p_0\\
&+\Big\{-(1-e(0))z^2+\{1+\sigma(1-e(0))\}zE(z)\\
&-E(z)\left[(1-\sigma)e(0)+\sigma E(z)\right]\Big\}p_1\\
&+e(0)\Big\{z^2-[\sigma z+(1-\sigma)]E(z)\Big\}p(A,2)\Big]\\
&/\left[z^3-z^2E(z)+\sigma(1-\sigma)zE(z)(E(z)-2)+\sigma(1-\sigma)E(z)^2\right] \quad . \qquad (7)
\end{aligned}
$$

Furthermore, we still have the balance equation for state 0:

$$
p_0=p_0e(0)+p_1e(0)+p(A,2)(1-\sigma)e(0)+p(B,2)\sigma e(0) \ ,
$$

which is equivalent with

$$
p(A,2)=\frac{[1-e(0)]p_0-p_1e(0)-p(B,2)\sigma e(0)}{(1-\sigma)e(0)} \ . \qquad (8)
$$

On account of (6)-(8), $U(z)$ is equal to

$$
\begin{aligned}
U(z)=\ & p_0+p_1z+Q_A(z)+Q_B(z)\\
=\ & (z-1)E(z)\Big[\Big\{(2\sigma-1)\sigma z^2e(0)+(1-\sigma)[(2\sigma+1)z^2-2\sigma(E(z)-1)z-\sigma E(z)]\Big\}p_0\\
&+\Big\{(2\sigma-1)\sigma z^2e(0)+(1-\sigma)\sigma z(2z-E(z))\Big\}p_1\\
&+(2\sigma-1)\sigma z^2e(0)p(B,2)\Big]\\
&/\left[z^3-z^2E(z)+\sigma(1-\sigma)zE(z)(E(z)-2)+\sigma(1-\sigma)E(z)^2\right] \ . \qquad (9)
\end{aligned}
$$

This equation still contains three unknown boundary probabilities p_0, p_1 and $p(B,2)$. It is, however, possible to express $p(B,2)$ as a function of p_0 and p_1 by invoking the "rate-in-rate-out" principle. The "rate-in-rate-out" principle expresses that in steady state the average

number of customers leaving the system in a slot equals the mean arrival rate λ, which leads to:

$$\lambda = p_1 + (2-\sigma)Q_A(1) + (1+\sigma)Q_B(1) \ . \tag{10}$$

Invoking (6) and (7) (for $z=1$) in (10) enables us to express $p(B,2)$ in terms of p_0 and p_1:

$$\begin{aligned} p(B,2) =& \Big[\big\{(1-\sigma)(1+\sigma) - e(0)\sigma(1-2\sigma)\big\}p_0 + \sigma\big\{1-\sigma-e(0)(1-2\sigma)\big\}p_1 \\ & +\sigma(1-\sigma) + \lambda\{1-2\sigma(1-\sigma)\} - 1\Big] \Big/ \Big[(1-2\sigma)\sigma e(0)\Big] \ . \end{aligned}$$

It can be shown that the same result can be obtained from the normalizing equation of the pgf $U(z)$, i.e., from the condition $U(1)=1$. Using this relation in (9) finally results into

$$\begin{aligned} U(z) =& (z-1)E(z)\Big[p_0\sigma(1-\sigma)\{z^2 + 2z(1-E(z)) - E(z)\} + p_1\sigma(1-\sigma)z\{z-E(z)\} \\ & - z^2\left\{\sigma(1-\sigma) - 1 + \lambda[1-2\sigma(1-\sigma)]\right\}\Big] \\ & \Big/\Big[z^3 - z^2E(z) + \sigma(1-\sigma)E(z)(E(z)-2)z + \sigma(1-\sigma)E(z)^2\Big] \ . \end{aligned} \tag{11}$$

The remaining boundary probabilities p_0 and p_1 can be determined by relying on the analytic property of pgf's, which implies that zeroes of the denominator of $U(z)$ inside the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ must also be zeroes of the numerator of $U(z)$.

It can be proved by using Rouché's theorem that the denominator of $U(z)$ has 3 zeroes inside the complex unit disk and that at least one zero equals 1. It is not difficult to see that the other two zeroes, say z_1 and z_2, are different from one (by taking the first derivative of the denominator, and invoking the stability condition and the analyticity of pgf's). When $z_1 \neq z_2$, the two remaining boundary probabilities p_0 and p_1 in (11) can now be calculated by solving the set of linear equations

$$U_N(z_i) = 0 \ , \quad i = 1,2 \ ,$$

where $U_N(z)$ is the numerator of $U(z)$ in equation (11). If $z_1 = z_2$, then the second equation reads

$$U_N^{'}(z_1) = 0 \ .$$

3.4 Performance measures

Once the two probabilities p_0 and p_1 have been computed, several performance measures of the system can be extracted from (11). First of all, the total mean system content, $\mathrm{E}[u]$ can be deduced by taking the first derivative of (11) at $z=1$:

$$\mathrm{E}[u] = \frac{2\sigma(1-\sigma)[(1-p_0)(3\lambda-2) + p_1(1-\lambda)] + 2\lambda(1-\lambda) + [\sigma^2 + (1-\sigma)^2]E^{''}(1)}{2[1-\lambda-\sigma(1-\sigma)(1-2\lambda)]} \ . \tag{12}$$

Owing to the independent occurrence of the two customer classes in the arrival process (class A with probability σ and class B with probability $1-\sigma$), we can also easily derive approximations for the per-class mean system contents, denoted as $\mathrm{E}[u_A]$ and $\mathrm{E}[u_B]$ for types A and B respectively:

$$\mathrm{E}[u_A] \approx \sigma \mathrm{E}[u] \quad ; \quad \mathrm{E}[u_B] \approx (1-\sigma)\mathrm{E}[u] \ . \tag{13}$$

The reasoning behind these approximations is that, possibly apart from the eldest customer in the system, all the (possible) other customers in the system have never belonged to the "set of the two eldest customers" in the system and their types have therefore not been influenced by the types of previous customers. This implies that these customers belong to class A and B with probabilities σ and $1-\sigma$ respectively, independently of all other customers. The type of the eldest customer, however, could be determined by the types of previous customers if the two eldest customers belonged to the same class one slot earlier. Upper and lower bounds for $\mathrm{E}[u_A]$ (or $\mathrm{E}[u_B]$) can be obtained by assuming that the eldest customer belongs to class A (or class B, respectively) with probability 1 or 0 respectively. This leads to

$$\sigma \mathrm{E}[u] - \sigma \leq \mathrm{E}[u_A] \leq \sigma \mathrm{E}[u] - \sigma + (1-p_0) \ ,$$

$$(1-\sigma)\mathrm{E}[u] - (1-\sigma) \leq \mathrm{E}[u_B] \leq (1-\sigma)\mathrm{E}[u] - (1-\sigma) + (1-p_0) \ ,$$

which shows that the approximations (13) are pretty good.

Next, using (the discrete-time version of) Little's law [11], i.e., $\mathrm{E}[u] = \lambda \mathrm{E}[d]$, we can derive the mean delay $\mathrm{E}[d]$ of an arbitrary customer from this as

$$\mathrm{E}[d] = \frac{2\sigma(1-\sigma)[(1-p_0)(3\lambda-2) + p_1(1-\lambda)] + 2\lambda(1-\lambda) + [\sigma^2 + (1-\sigma)^2]E''(1)}{2\lambda[1-\lambda-\sigma(1-\sigma)(1-2\lambda)]} \ . \tag{14}$$

Approximations for the per-class mean delays $\mathrm{E}[d_A]$ and $\mathrm{E}[d_B]$ can be analogously obtained by applying Little's law to the class-A or class-B customers individually, which leads to

$$\mathrm{E}[d_A] = \frac{\mathrm{E}[u_A]}{\sigma\lambda} \approx \frac{\mathrm{E}[u]}{\lambda} = \mathrm{E}[d] \tag{15}$$

and

$$\mathrm{E}[d_B] = \frac{\mathrm{E}[u_B]}{(1-\sigma)\lambda} \approx \frac{\mathrm{E}[u]}{\lambda} = \mathrm{E}[d] \ .$$

The above result shows that the mean delay of an arbitrary type-A customer is approximately equal to the mean delay of an arbitrary type-B customer, which is intuitively acceptable in view of the global FCFS service policy which does not discriminate between the two customer classes.

4 Individual queues

In this section, we summarize results for an analogous model as described in section 2, but without the global FCFS policy, implying that customers cannot be blocked by customers of other types. This thus corresponds to a system with two individual single-server queues with single-slot service times, one with mean arrival rate $\sigma\lambda$ and the other with mean arrival rate $(1-\sigma)\lambda$.

4.1 Stability

It is not difficult to see that the stability condition for this system is given by

$$\lambda < \min\left\{\frac{1}{\sigma}, \frac{1}{1-\sigma}\right\} \ . \tag{16}$$

This should be compared with the inequality (1) in case of global FCFS. It is straightforward to prove that

$$\frac{1-\sigma(1-\sigma)}{1-2\sigma(1-\sigma)} \leq \min\left(\frac{1}{\sigma}, \frac{1}{1-\sigma}\right) \ , \tag{17}$$

for $0 < \sigma < 1$. Hence, in general, the stability condition is more stringent in case of global FCFS, meaning that the maximum tolerable arrival rate is smaller. When $\sigma = 0$ or $\sigma = 1$, both systems are equivalent with a single-server queue fed by an arrival process with mean arrival rate λ and stability condition $\lambda < 1$. It is worth noting that the inequality (17) also implies that the stability condition (1) in case of global FCFS not only guarantees global stability (for the total system content), but also individual stability for each customer type.

4.2 Performance measures

The pgf of the system content at random slot boundaries in one individual queue is given by the well-known formula for a discrete-time single-server system with service times of one slot (see e.g. [3], [21]):

$$U_i(z) = \frac{\left(1 - E_i^{'}(1)\right)(z-1)E_i(z)}{z - E_i(z)} \ , \quad i = A, B \ ,$$

where $U_i(z)$ is the pgf of the system content in the class-i queue. The corresponding mean system content of type i is given by

$$\mathrm{E}\left[u_i\right] = U_i^{'}(1) = E_i^{'}(1) + \frac{E_i^{''}(1)}{2\left[1 - E_i^{'}(1)\right]} \ .$$

Here the function $E_i(z)$ $(i = A, B)$ denotes the pgf of the number of type-i arrivals per slot. In view of the lack of interclass correlation in the arrival process, $E_A(z)$ and $E_B(z)$ are given by

$$E_A(z) = E(1 - \sigma + \sigma z) \ ,$$

$$E_B(z) = E(\sigma + (1-\sigma)z) \ ,$$

which leads to the following explicit expressions for $\mathrm{E}\left[u_A\right]$ and $\mathrm{E}\left[u_B\right]$:

$$\mathrm{E}\left[u_A\right] = \sigma\lambda + \frac{\sigma^2 E''(1)}{2(1-\sigma\lambda)} \ ,$$

$$\mathrm{E}\left[u_B\right] = (1-\sigma)\lambda + \frac{(1-\sigma)^2 E''(1)}{2(1-(1-\sigma)\lambda)} \ .$$

The system contents in buffers A and B are, in general, not independent. (A notorious exception is the case of Poisson arrivals, where $E(z) = e^{\lambda(z-1)}$, because in this particular case, the two partial arrival streams for the two customer classes happen to be independent.) As a result, the pgf of the total system content is (in general) not just the product of the individual pgf's. However, the mean value of the total system content ($\mathrm{E}[u]$) is always given by the sum of the mean values of the system contents in the individual queues, leading to

$$\begin{aligned} \mathrm{E}[u] = &\ \mathrm{E}[u_A] + \mathrm{E}[u_B] \\ = &\ \lambda + E''(1)\left(\frac{\sigma^2}{2(1-\sigma\lambda)} + \frac{(1-\sigma)^2}{2[1-(1-\sigma)\lambda]}\right) \quad . \end{aligned} \tag{18}$$

Again, we can derive the global mean customer delay $\mathrm{E}[d]$ from this by means of (the discrete-time version of) Little's law [11]:

$$\mathrm{E}[d] = \frac{\mathrm{E}[u]}{\lambda} = \ 1 + E''(1)\left(\frac{\sigma^2}{2\lambda(1-\sigma\lambda)} + \frac{(1-\sigma)^2}{2\lambda[1-(1-\sigma)\lambda]}\right) \quad .$$

Similarly, the per-class mean delays $\mathrm{E}[d_A]$ and $\mathrm{E}[d_B]$ are obtained as

$$\mathrm{E}[d_A] = \frac{\mathrm{E}[u_A]}{\sigma\lambda} = 1 + \frac{\sigma E''(1)}{2\lambda(1-\sigma\lambda)} \tag{19}$$

and

$$\mathrm{E}[d_B] = \frac{\mathrm{E}[u_B]}{(1-\sigma)\lambda} = 1 + \frac{(1-\sigma)E''(1)}{2\lambda[1-(1-\sigma)\lambda]} \quad .$$

Note that the global mean delay $\mathrm{E}[d]$ can also be derived from

$$\mathrm{E}[d] = \sigma\mathrm{E}[d_A] + (1-\sigma)\mathrm{E}[d_B] \quad .$$

5 Influence of global FCFS and relative load distribution

In this section, we investigate the influence of the global FCFS policy and the relative load distribution on the behavior of the system. We have therefore first depicted the maximum tolerable arrival rate λ, as derived from equations (1) and (16), versus the load-distribution parameter σ in Fig. 2. On the other hand, the total mean system content $\mathrm{E}[u]$, as defined by equations (12) and (18), is shown versus λ and versus σ in Figs. 3 and 4 respectively. Curves are plotted for the global FCFS policy as well as for individual queues (i.e., without global FCFS). In Figs. 3 and 4, we have assumed that the number of arrivals per slot has a geometric distribution with mean λ.

We observe from Fig. 3 that global FCFS only has a minor impact on the mean system content when $\lambda < 1$. When $\lambda < 1$, fewer customers compete for service, so that an arriving customer more probably enters a sparsely populated system, and thus experiences nearly no blocking of customers of the other type. On the other hand, when $\lambda > 1$, an arriving customer is more likely to arrive in a densely populated system and is thus more likely to be hindered by

customers of the other type. This leads to a larger mean system content and a smaller range of tolerable combinations of λ and σ (i.e., those combinations where the load remains smaller than one) as compared to two individual queues. Note, for instance, that when $\sigma = 0.5$, the maximum tolerable arrival rate is 1.5 in case of global FCFS instead of 2 in case of individual queues (Fig. 2). Indeed, the probability that the second eldest customer is of the same type as the eldest equals 0.5, in which case only one server processes, whereas with probability 0.5 the two eldest customers are of distinct types so that both servers are active. Hence, the mean number of active servers per slot in a nearly saturated system with global FCFS and $\sigma = 0.5$ equals $1 \cdot 0.5 + 2 \cdot 0.5 = 1.5$.

The figures further illustrate that the system, regardless of whether the global FCFS policy is adopted or not, performs best when $\sigma = 0.5$, in terms of smaller $\mathrm{E}[u]$ and larger maximum tolerable arrival rate. The reason is that both servers process an equal fraction of the customers, i.e., work is spread fairly amongst the servers. Distributing the work in computer networking is called "load balancing" and our result here is in agreement with the knowledge that load balancing leads to a more efficient use of resources, a better throughput, et cetera [7, 9, 18]. Figs. 2 and 3 further show that, regardless of the policy, the system performs worst when $\sigma = 0$ (or, $\sigma = 1$). In these cases, all customers have to be processed by the same server, whereas the other server is superfluous. As a result, the system degrades to a single-server system, with stability condition $\lambda < 1$, which is reflected clearly in Figs. 2 and 3.

Note that we have only shown values of $\sigma \leq 0.5$ in Fig. 3, as $\sigma = \alpha$ and $\sigma = 1 - \alpha$ ($0 \leq \alpha \leq 1$) lead to the same results. The key observation to understand the latter is that for the operation of the system only equality or non equality of the two customer types is of importance. In fact, a system with $\sigma = 1 - \alpha$ can be conceived as a system with $\sigma = \alpha$ whereby the names of the types A and B have been "swapped". There thus exists a kind of symmetry in the customer types around the value $\sigma = 0.5$. This symmetry can be observed clearly in Figs.

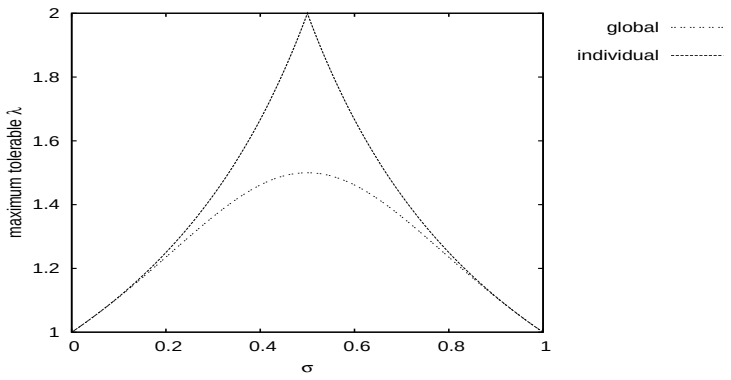

Figure 2: Maximum tolerable arrival rate λ versus the load-distribution parameter σ

2 and 4: the curves are symmetric around their best case $\sigma = 0.5$ and the more σ differs from 0.5, the worse the system behaves. Indeed, the more σ differs from 0.5, the more the system becomes similar to a single-server system. As σ deviates more from 0.5, the stability condition is eventually violated for $\lambda = 1.4$ in Fig. 4, which explains the vertical asymptotes in that case.

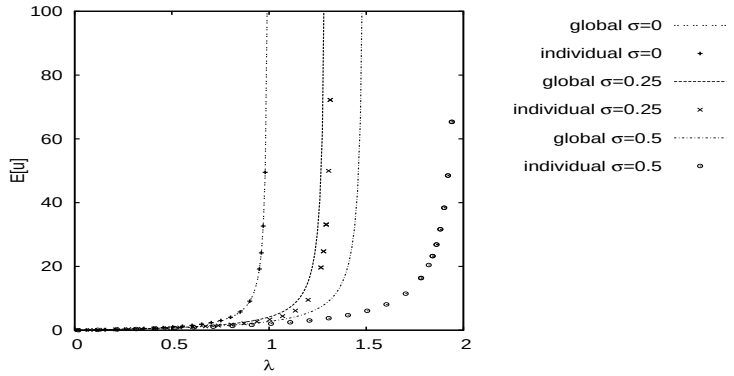

Figure 3: Mean system content $\mathrm{E}[u]$ versus the mean arrival rate λ, for various values of the load-distribution parameter σ

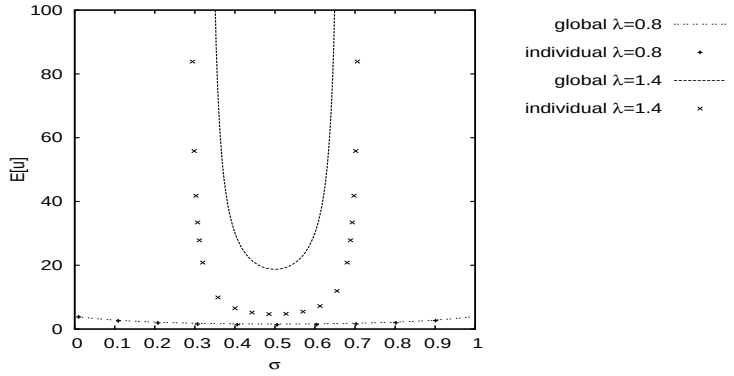

Figure 4: Mean system content $\mathrm{E}[u]$ versus the load-distribution parameter σ, for various values of the mean arrival rate λ

6 Conclusions and further research

In this paper, we have analyzed a discrete-time queueing model where two types (classes) of customers, both to be served by their own dedicated server, are accommodated in one common FCFS queue (*global FCFS*). A customer belongs to the first class with probability σ and to the second class with probability $1-\sigma$, independently from customer to customer. We have deduced the stability condition and have calculated the pgf of the total system content, whereafter we have compared these results with those whereby individual queues are provided for the two servers. We have demonstrated that when the mean total arrival rate (λ) is smaller than 1, global FCFS only has a minor impact on the total mean system content, whereas the opposite holds when $\lambda > 1$. We have also shown that the system performance is symmetric around $\sigma = 0.5$.

There are a number of possible extensions to this work. First, the independence assumption of the types of consecutive customers in the arrival stream could be relaxed. The simplest possible extension in this respect would probably be to assume that the types of consecutive customers form a first-order Markov chain. This comes down to assuming that the probability that the next customer belongs to class A or B depends on the type of the previous customer. In fact, a very specific special case of this kind of model was considered in our earlier paper

[4], where we introduced a "cluster parameter" in the description of the arrival process, which denotes the probability that the next customer has the *same type* as the previous customer. In [4], however, we assumed that the cluster parameter did not depend on the type of the previous customer, which basically comes down to assuming equal loads for both customer classes. This could be relaxed to two class-dependent cluster parameters, i.e., arbitrary transition probabilities for the Markov chain mentioned above. However, it is to be expected that the analysis of this more general case would be considerably more complicated than the analyses in [4] and in the current paper. Of course, even more general assumptions than first-order Markov could be envisaged, such as alternating periods (of random length) in which only customers of type A or B respectively, arrive in the system, and so on. Another restriction of the current work is the assumption that all service times are deterministically equal to 1 slot. Although this assumption greatly simplifies the analysis of the model, it does imply that customers can never "overtake" each other while being served. If the service times, however, were random (and, hence variable), the latter phenomenon could occur and possibly affect the blocking in the system. We plan to tackle such generalizations of the model in future research.

Acknowledgment This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

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

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

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

[4] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. Impact of blocking when customers of different classes are accommodated in one common queue. In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems (ICORES)*, Villamoura, Portugal, February 2012.

[5] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. Influence of relative traffic distribution in nodes with blocking: an analytical model. In *Proceedings of the European Simulation and Modelling Conference (ESM)*, pages 136–143, Essen, Germany, October 2012.

[6] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *European Journal of Operational Research*, 223:123–132, 2012.

[7] V. Cardellini, M. Colajanni, and P.S. Yu. Dynamic load balancing on web-server systems. *IEEE Internet Computing*, 3(3):28–39, 1999.

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

[9] T.C.K. Chou and J.A. Abraham. Load balancing in distributed systems. *IEEE Transactions on Software Engineering*, 8(4):401–412, 1982.

[10] S. De Vuyst, S. Wittevrongel, and H. Bruneel. Place reservation: Delay analysis of a novel scheduling mechanism. *Computers & Operations Research*, 35(8):2447–2462, 2008.

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

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

[13] A. Kesselman, K. Kogan, and M. Segal. Improved competitive performance bounds for CIOQ switches. *Algorithmica*, 63(1-2):411–424, 2012.

[14] W. Mélange, H. Bruneel, B. Steyaert, and J. Walraevens. A two-class continuous-time queueing model with dedicated servers and global FCFS service discipline. In *Proceedings of the 18th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pages 14–27, Venice, Italy, June 2011.

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

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

[17] I. Rubin and Z. Tsai. Message delay analysis of multiclass priority TDMA, FDMA, and discrete-time queueing systems. *IEEE Transactions on Information Theory*, 35(2):637–647, 1989.

[18] R. Schoonderwoerd, J.L. Bruten, O.E. Holland, and L.J.M. Rothkrantz. Ant-based load balancing in telecommunications networks. *Adaptive Behavior*, 5(2):169–207, 1996.

[19] D. Shah, J.N. Tsitsiklis, and Y. Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3-4):375–384, 2011.

[20] D.R. Smith and W. Whitt. Resource sharing for efficiency in traffic systems. *The Bell System Technical Journal*, 60(1):39–55, 1981.

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

[22] I.M. Verloop, U. Ayesta, and S. Borst. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems - Theory and Applications*, 20(4):473–509, 2010.

[23] J. Walraevens, D. Fiems, and H. Bruneel. Time-dependent performance analysis of a discrete-time priority queue. *Performance Evaluation*, 65:641–652, 2008.

[24] H. Yu, S. Ruepp, and M.S. Berger. Enhanced first-in-first-out-based round-robin multicast scheduling algorithm for input-queued switches. *IET Communications*, 5(8):1163–1171, 2011.