Stability Analysis of Global FCFS and Presorting Service Discipline

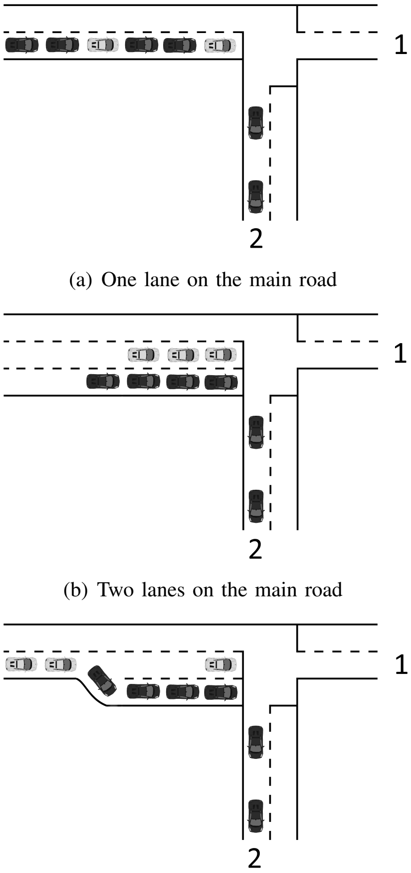
Willem Mélange, Joris Walraevens, Dieter Claeys, Bart Steyaert and Herwig Bruneel
Department of Telecommunications and Information Processing
Ghent University - UGent
E-mail: {wmelange,jw,dc,bs,hb}@telin.UGent.be

***Abstract*—In this paper, we consider a continuous-time queueing system with two different types (1 and 2) of customers with two dedicated servers (also named 1 and 2). This means server 1 (2) can only serve customers of type 1 (2). The goal of this paper is to determine the stability condition for our system with global first-come-first-serve (FCFS) and presorting service discipline, i.e., all arriving customers are accommodated in one single FCFS queue, regardless of their types, with an exception of the first N customers. For the first N customers the FCFS rule holds only within the types, i.e., customers of different types can overtake each other in order to be served. The motivation for our work comes from traffic and is to be able to give advise about the optimal length of filter lanes, i.e., lanes reserved for vehicles making a specific turn at a junction.**

***Keywords*—queueing, stability, blocking, global FCFS, presorting.**

I. Introduction

The motivation for this work is an every day problem in traffic. Traffic jams might occur for multiple reasons. One reason are traffic junctions. Consider, for instance, the following situation: vehicles approach a junction with two possible destinations (1 and 2) as seen in Fig. 1. In traffic context, it is often not physically feasible to provide two separate lanes for each possible destination (as seen in Fig. 1(b)). If it would be, vehicles for both directions can be kept apart completely. The other extreme occurs much more frequently, namely, when there is one lane on the main road (Fig. 1(a)). In the case where there is only one lane on the main road, it is possible that vehicles on that road heading for destination 1 may be hindered or even blocked by vehicles heading for destination 2, even when the subroad leading to destination 1 is free, simply because cars that go to 2 are in front of them. In other words, there is a first-come-first-serve (FCFS) order on the main road regardless what destination they have. In the rest of this paper we will call this service discipline global FCFS (gFCFS). In queueing theory terms, a service discipline where there are 2 types of customers that are accommodated in a single queue and who are served in a FCFS manner regardless of their type. When we look at the case in Fig. 1(a), there is even a global first-in-first-out (gFIFO) order on the main road. At any given time, at most one server will be working. A possible way to minimize the impact of this blocking phenomenon is the use of filter lanes, i.e., lanes reserved for vehicles making a specific turn at a junction (as seen in Fig. 1(c)). It is clear that in this case, we cannot longer talk about gFCFS or gFIFO as service discipline. In the rest of the paper we will call this new service discipline, which can be seen as sort of relaxation of the gFCFS service discipline, gFCFS with presorting (P-gFCFS). Again in queueing theory terms, a service discipline where there are 2 types of customers that are accommodated in a single queue and who are served in a FCFS manner regardless of their type with an exception for the first N customers. Thus, the customers can only be served if they are one of the first N customers in the system and there are no customers of the same type in front of them. Fig. 1(c) is an example of such a system with a P-gFCFS service discipline. In this specific case, N is equal to 4. If the vehicle is in the first 4 vehicles on the main road and there is no vehicle in front of this vehicle with the same direction, the vehicle will be able to drive without delay caused by other vehicles to the destination that vehicle desires.



(a) One lane on the main road

(b) Two lanes on the main road

(c) One lane on the main road with filter lane

Fig. 1. Light grey vehicles with destination 1 and dark grey vehicles with destination 2 approaching a traffic junction

ISBN: 978-1-61208-283-7

However, if there are 4 vehicles with another destination in front of this vehicle, even when there are no vehicles in front of this vehicle with the same destination, the vehicle is not able to drive without delay caused by other vehicles to his destination (as seen in Fig. 1(c)). We refer to [1], [2] for a general overview and validation of modelling traffic flows with queueing models.

Analogously, at a security checkpoint (e.g., at an international airport or train station) people are usually body-searched by someone of the same gender. As a result, when a group of friends of the same gender arrive, the people of the opposite gender behind them may have to wait until the whole group has been checked, even when the other security person is available, at least when it is not allowed to overtake at the security checkpoint (which is often the case for security reasons). Here, we can also have an relaxation of the gFCFS service discipline. The security person can hand-pick a person from the waiting line to come to the front of the line. We also presume that the security person will only do this for one of the first N persons in the waiting line.

In [3] and [4], we already got some insight in the impact of the blocking phenomenon caused by a gFCFS service discipline on the performance of the involved systems (or a P-gFCFS with parameter N equal to 2). As stated earlier, in this paper, we want to relax the gFCFS rule and get some insight in the impact of this relaxation. In other words, we want to investigate the impact of the N-gFCFS service discipline on the performance of the involved systems.

The structure of the rest of the paper is as follows: we first tackle in Section II the (more simple) problem where the types in the arrival stream of customers are independent. In this Section, first the mathematical model is described in Subsection II-A. Next we analyse the stability condition in Subsection II-B. Then we tackle in Section III the (more difficult) problem where the types in the arrival stream of customers are dependent. The same structure is used as in the Section II where the types of customers in the arrival stream are independent. The paper continuous with a discussion about the results and some numerical examples in Section IV. Finally, some conclusions are drawn and future research is suggested in Section V.

II. Uncorrelated Types in the Arrivals

We start with the case of uncorrelated types in the arrivals. This is for instance an adequate model for traffic junctions, as the destination of consecutive cars can largely be regarded as independent.

A. Mathematical Model

We consider a continuous-time queueing model with infinite waiting room. There are two servers, where server 1 is working at rate μ_1 and server 2 at rate μ_2 (exponential service times). There are two types (classes) of customers. Each of the two servers is dedicated to a given class of customers. In this case, server 1 always serves customers of type 1 and server 2 always serves customers of type 2. The customers are served as follows: if both types are present in the first N customers in the system, the first customer of each type can be served by its server. The customers not in the first N customers are served in a global FCFS way, i.e., they are blocked not only by customers of their own type in front of them, but also by customers of different type. We will call this service discipline global FCFS with presorting (P-gFCFS). For example, if the first $i-1$ customers are of type 1 and the i-th customer is of type 2, then this customer can be served by server 2 if $i \leq N$. However, if the first N customers are of type 1 and the $N+1$-th customer is of type 2, then this customer cannot be served by server 2 even if the server is idle.

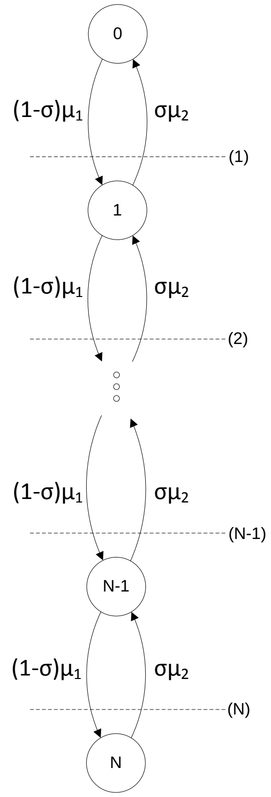


Fig. 2. $(N+1)$-state Markov chain to determine the stability condition of the system with P-gFCFS with uncorrelated types in the arrival stream

The customers enter the system according to a Poisson arrival process with mean arrival rate λ. With probability σ, the customer is of type 1 and with probability $1-\sigma$ the customer is of type 2.

B. Analysis of the Stability Condition

When looking at the stability condition, we can presume that the system is constantly provided with new customers and the system will therefore be filled with at least N customers all the time. Note that we are only interested in the number of customers of type 1 and 2 in the first N customers of the system. Thus, the exact queueing order of the types of the first N customers is of no importance. These observations lead to the $(N+1)$-state Markov chain in Fig. 2. The state m, m customers of the first N customers are of type 2 (and thus $N-m$ of type 1). The rate to go from state m to state $m-1$, is $\sigma\mu_2$; namely a rate μ_2 to end the service in state m of the customer with type 2 multiplied with the probability σ that the new N-th customer of our system is of type 1. Similarly, the rate to go from state m to state $m+1$, is $(1-\sigma)\mu_1$.

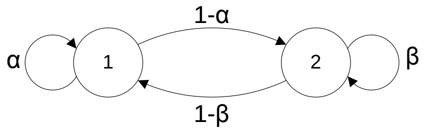


Fig. 3. 2-state Markov chain to determine the type of an arriving customer

Fig. 2 models the well-known birth-and-death process for a $M|M|1|N$ queue [5] and the probability to be in state m is known to be given by

$$p(m) = \frac{\left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^m \left(1 - \frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)}{1 - \left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^{N+1}}. \tag{1}$$

Having obtained the $p(m)$'s, we can now move to the stability condition. Therefore we postulate that the average amount of work per unit time that enters the system (γ) is smaller than the average amount of work the system can execute per unit time, i.e., the average amount of work the system would execute per unit time when it would be constantly provided with new customers. Here, the system is able to execute 2 units of work per unit of time when both servers are able to work (when the system is in one of the states 1 to $N-1$). The system is able to execute 1 unit of work per unit of time when only one server is able to work (when the system is in state 0 or N). The stability condition is thus

$$\gamma < p(0) + 2\sum_{m=1}^{N-1} p(m) + p(N) \tag{2}$$

$$\gamma < \frac{\left(1 + \frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)\left(1 - \left(\frac{1-\sigma)\mu_1}{\sigma\mu_2}\right)^N\right)}{1 - \left(\frac{(1-\sigma)\mu_1}{\sigma\mu_2}\right)^{N+1}}. \tag{3}$$

where γ (the average amount of work that enters the system per unit time) is given by

$$\gamma \triangleq \lambda\left(\frac{\sigma}{\mu_1} + \frac{1-\sigma}{\mu_2}\right). \tag{4}$$

Equation (3) can then be written as

$$\lambda < \frac{\left(\frac{\sigma}{\mu_1}\right)^N - \left(\frac{1-\sigma}{\mu_2}\right)^N}{\left(\frac{\sigma}{\mu_1}\right)^{N+1} - \left(\frac{1-\sigma}{\mu_2}\right)^{N+1}}. \tag{5}$$

which says that on average, there are not more arrivals than service completions.

III. Correlated Types in the Arrivals

We now turn to the case that some correlation in the types of consecutively arriving customers is present. This can, for instance, be the case in the modelling of a security check point, where partners of different sex (negative correlation) may arrive more frequently, or groups of people, all of the same sex (positive correlation).

A. Mathematical Model

The model is the same as in Section II-A except for the arrival stream. Now the customers enter the system according to a Poisson arrival process with mean arrival rate λ. The type of the arriving customer is determined by a two-state Markov chain (see Fig. 3). If the previous customer is of type 1, then the customer is of type 1 with probability α and of type 2 with probability $(1-\alpha)$. If the previous customer is of type 2, then the customer is of type 1 with probability $(1-\beta)$ and of type 2 with probability β. Notice here already that we can transform α and β, in two other parameters σ and K that have a more intuitive meaning. The transformations from (α,β) to (σ,K) are

$$\sigma = \frac{1-\beta}{2-\alpha-\beta}, \tag{6}$$

$$K = \frac{1}{2-\alpha-\beta} \tag{7}$$

and from (σ,K) to (α,β) are

$$\alpha = 1 - \frac{1-\sigma}{K}, \tag{8}$$

$$\beta = 1 - \frac{\sigma}{K}. \tag{9}$$

The intuitive meaning behind the parameter σ is the given relative frequency distribution of the type of the customers. The fraction of customers that are of type 1 (2) is σ ($1-\sigma$ respectively). The parameter K on the other hand gives a clear indication about the correlation. The parameter is directly proportional to the mean number of customers of the same type that arrive back-to-back. More specifically, we have

$$E\left[\text{number of customers of type 1 arriving back-to-back}\right] = \frac{1}{1-\beta} = \frac{K}{\sigma}, \tag{10}$$

$$E\left[\text{number of customers of type 2 arriving back-to-back}\right] = \frac{1}{1-\alpha} = \frac{K}{1-\sigma}, \tag{11}$$

where $E\left[\cdots\right]$ stands for the the expected value of what's between brackets. Notice here that when K equals 1, the types of customers in the arrival stream are uncorrelated, and the model transforms to that of Section II.

B. Analysis of the Stability Condition

We take the same approach as in Section II-B. The Markov chain corresponding to Fig. 2 is more complicated in this case. Now we do not only have to keep track of the number of type 1 and 2 customers in the first N customers, but also of the type of the "last" customer in this set of customers. These observations lead to the $2N$-state Markov chain in Fig. 4, where in state (m,t), m customers of the N first customers are of type 2 (and thus $N-m$ of type 1) and the "last" customer is of type t. Notice that we do not have states $(0,2)$ and $(N,1)$ since the "last" customer cannot be of type 2 (1) if all N customers are of type 1 (2).

The balance equations (see transitions through the dotted lines ($1b$) to ($(N-1)b$) in Fig. 4) are

$$\mu_2 p(m,1) = \mu_1 p(m,2), \tag{12}$$

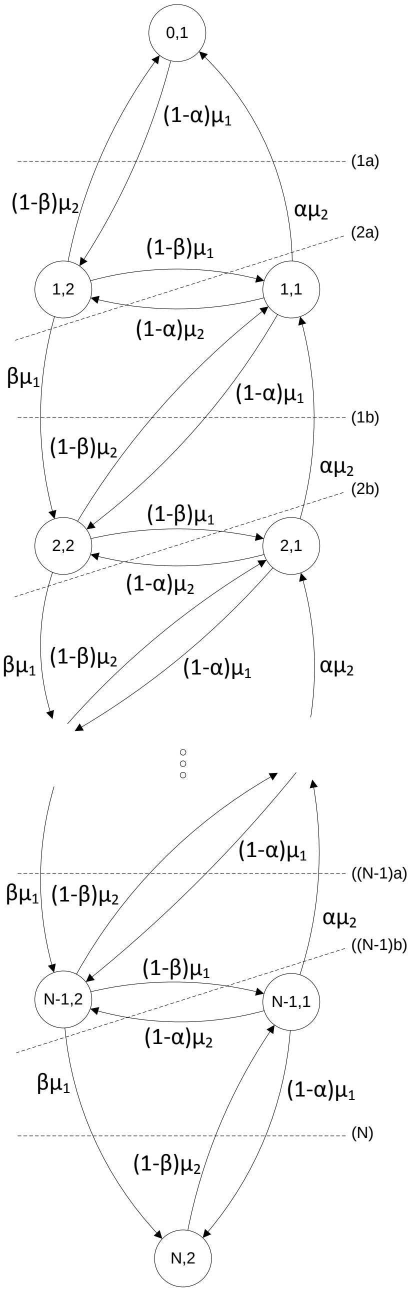


Fig. 4. $2n$-state Markov chain to determine the stability condition of the system with P-gFCFS with correlated types in the arrival stream

where $m = 1, \cdots, N-1$ and (see transitions through the dotted lines $(2a)$ to $((N-1)a)$ in Fig. 4)

$$\begin{aligned}(1-\beta)\mu_2 p(m,2) + \alpha\mu_2 p(m,1)& \\ = (1-\alpha)\mu_1 p(m-1,1) + \beta\mu_1 p(m-1,2),& \quad (13)\end{aligned}$$

where $m = 2, \cdots, N-1$. Equations (12) and (13) can be rewritten as

$$p(m,1) = \frac{\mu_1}{\mu_2} p(m,2), \quad (14)$$

$$\begin{aligned}((1-\beta)\mu_2 + \alpha\mu_1)p(m,2)& \\ = \left((1-\alpha)\frac{\mu_1^2}{\mu_2} + \beta\mu_1\right) p(m-1,2)& \quad (15)\end{aligned}$$

or even further as

$$p(m,1) = \frac{\mu_1}{\mu_2} p(m,2), \quad (16)$$

$$p(m,2) = \frac{(1-\alpha)\mu_1^2 + \beta\mu_1\mu_2}{(1-\beta)\mu_2^2 + \alpha\mu_1\mu_2} p(m-1,2). \quad (17)$$

This yields for $m = 1, \cdots, N-1$

$$p(m,1) = \frac{\mu_1}{\mu_2} \theta_1^{m-1} p(1,2) \quad (18)$$

$$p(m,2) = \theta_1^{m-1} p(1,2), \quad (19)$$

where

$$\theta_1 = \frac{(1-\alpha)\mu_1^2 + \beta\mu_1\mu_2}{(1-\beta)\mu_2^2 + \alpha\mu_1\mu_2}. \quad (20)$$

The balance equation corresponding with transition $(1a)$ reads

$$(1-\alpha)\mu_1 p(0,1) = (1-\beta)\mu_2 p(1,2) + \alpha\mu_2 p(1,1) \quad (21)$$

and using (12)

$$p(1,2) = \frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2 + \alpha\mu_1} p(0,1). \quad (22)$$

Using (22) in (18) and (19), we find

$$p(m,1) = \frac{\mu_1}{\mu_2} \theta_1^{m-1} \frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2 + \alpha\mu_1} p(0,1), \quad (23)$$

$$p(m,2) = \theta_1^{m-1} \frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2 + \alpha\mu_1} p(0,1). \quad (24)$$

The last balance equation corresponding with transition (N) leads to

$$p(N,2) = \frac{(1-\alpha)\mu_1^2 + \beta\mu_1\mu_2}{(1-\beta)\mu_2^2} p(N-1,2), \quad (25)$$

where we used (12) to eliminate $p(N-1,1)$. The normalization condition

$$\sum_{m=0}^{N} (p(m,1) + p(m,2)) = 1, \quad (26)$$

where $p(0,2) = p(N,1) = 0$ by definition, finally yields $p(0,1)$. Just as in Section II-B we can write the stability condition as the inequality that the average amount of work per unit time that enters the system (γ) is smaller than the average amount of work the system can execute per unit time. In this case, the stability condition is given by

$$\gamma < p(0,1) + 2\sum_{m=1}^{N-1} (p(m,1) + p(m,2)) + p(N,2) \quad (27)$$

$$\gamma < \frac{(1 + a\theta_1^{N-1})(\theta_1 - 1) + 2b(\theta_1^{N-1} - 1)}{(1 + a\theta_1^{N-1})(\theta_1 - 1) + b(\theta_1^{N-1} - 1)} \quad (28)$$

with

$$a = \frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2}, \quad (29)$$

$$b = \left(\frac{(1-\alpha)\mu_1}{(1-\beta)\mu_2 + \alpha\mu_1}\right)\left(\frac{\mu_1}{\mu_2} + 1\right). \quad (30)$$

To make the results more intuitive we use the transformations from equations (8) and (9). Here, we also already see that not

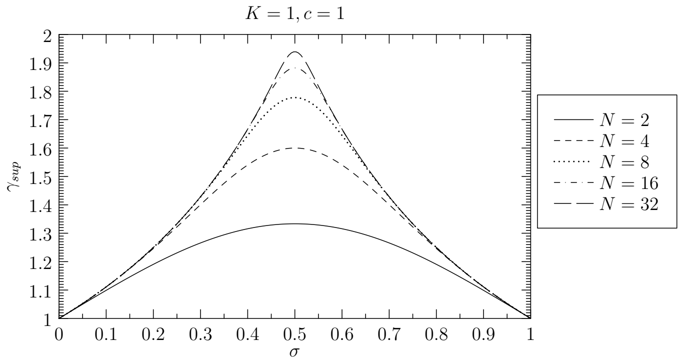


Fig. 5. γ_{sup}, least upper bound of the set of γ values where the system is stable, versus parameter σ with $K = 1$ and $c = 1$

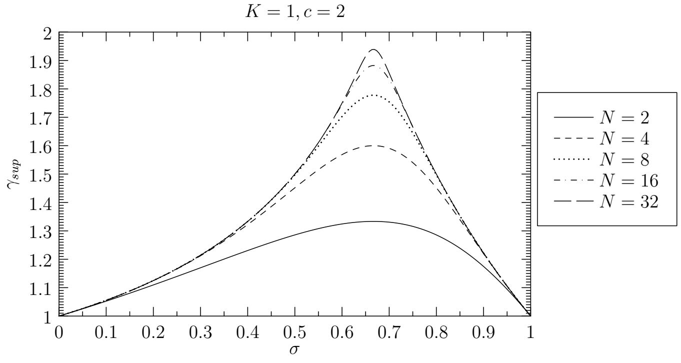


Fig. 6. γ_{sup}, least upper bound of the set of γ values where the system is stable, versus parameter σ with $K = 1$ and $c = 2$

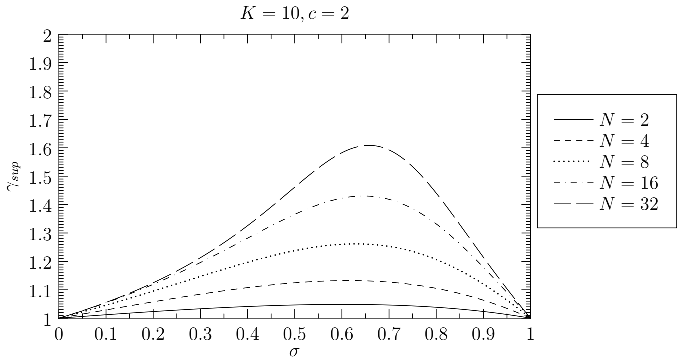


Fig. 7. γ_{sup}, least upper bound of the set of γ values where the system is stable, versus parameter σ with $K = 10$ and $c = 2$

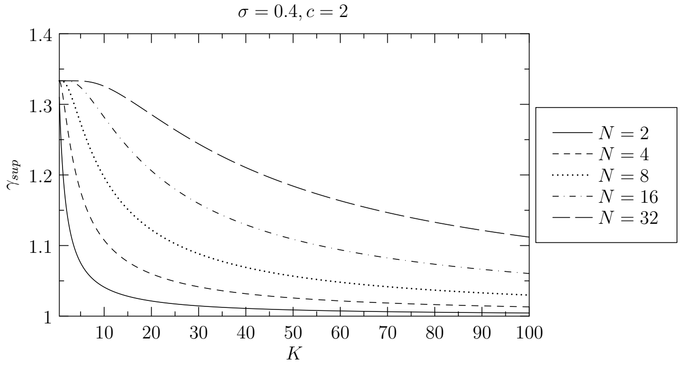


Fig. 8. γ_{sup}, least upper bound of the set of γ values where the system is stable, versus parameter K with $\sigma = 0.4$ and $c = 2$

the exact values of μ_1 and μ_2 are of importance but only the ratio. The stability condition becomes

$$\gamma < \frac{(1 + a\theta_1^{N-1})(\theta_1 - 1) + 2b(\theta_1^{N-1} - 1)}{(1 + a\theta_1^{N-1})(\theta_1 - 1) + b(\theta_1^{N-1} - 1)} \tag{31}$$

with

$$c = \frac{\mu_1}{\mu_2}, \tag{32}$$

$$\theta_1 = c\frac{K + (1-\sigma)c - \sigma}{cK - (1-\sigma)c + \sigma}, \tag{33}$$

$$a = c\frac{(1-\sigma)}{\sigma}, \tag{34}$$

$$b = \left(\frac{(1-\sigma)}{cK - (1-\sigma)c + \sigma}\right)(c+1)\,. \tag{35}$$

Notice that when $\theta_1 = 1$, we need to use l'Hôpital's rule to determine the value of the right hand side of (31). Thus, when $\theta_1 = 1$, the stability condition is given by

$$\gamma < \frac{1 + aN\theta_1^{N-1} - (N-1)(a-2b)\theta_1^{N-1}}{1 + aN\theta_1^{N-1} - (N-1)(a-b)\theta_1^{N-1}}. \tag{36}$$

IV. Discussion of the Results and Numerical Examples

It is always interesting to look at the extreme situations. The first one is when $N = 1$. This means that we have a first-in-first-out queue regardless of the types of customers. The stability condition (31) becomes

$$\gamma < 1 \tag{37}$$

This is what we would also intuitively expect. The amount of work the system can execute per unit time is 1 (right-hand side of equation (37)). In other words, at any given time only one server is working. The second extreme situation is when $N = \infty$. In this case, seperate queues for each server are present. We have to split up this situation in three cases. The first case is when $\theta_1 = 1$, rewritten as $\frac{\sigma}{\mu_1} = \frac{1-\sigma}{\mu_2}$ (independent of K) or the load is balanced between both servers. The stability condition (31) becomes

$$\gamma < 2 \tag{38}$$

This is, again, what we also would expect intuitively. The amount of work the system can execute per unit time is 2 (right-hand side of equation (42)). In other words, at any given time both servers are working provided the system is constantly provided with new customers. Notice that this is only possible when the load is balanced. The second case is when $\theta_1 > 1$, rewritten as $\frac{\sigma}{\mu_1} < \frac{1-\sigma}{\mu_2}$ or server 1 has a heavier load. The stability condition (31) becomes

$$\gamma < \frac{a\theta_1 - a + 2b}{a\theta_1 - a + b} \tag{39}$$

or further simplified as

$$(1-\sigma)\lambda < \mu_2. \tag{40}$$

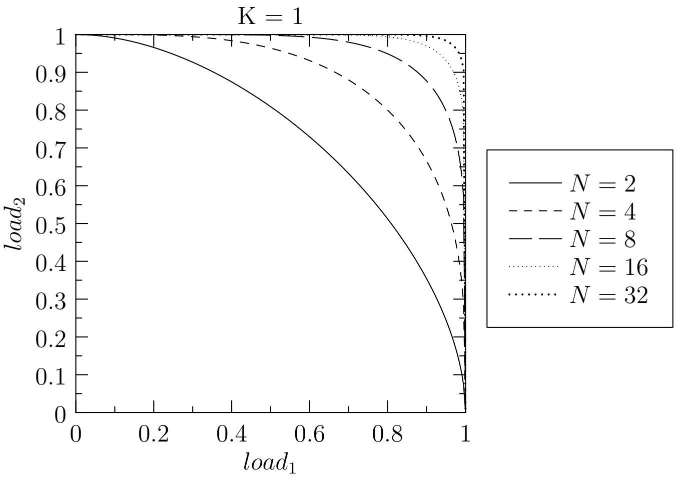


Fig. 9. $load_2$ versus $load_1$ with uncorrelated types of customers in the arrival stream ($K = 1$)

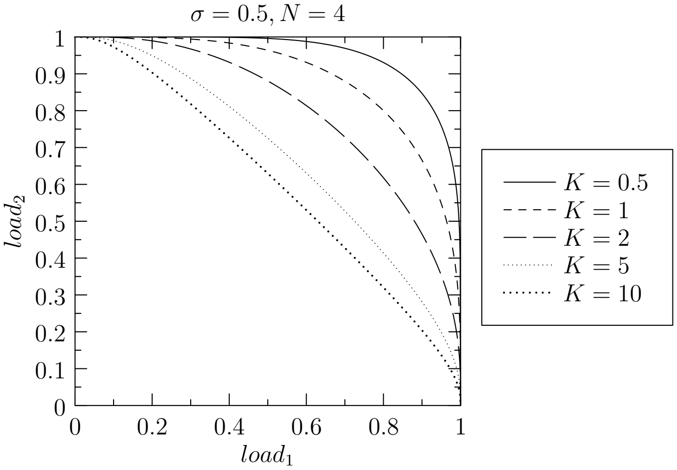


Fig. 10. $load_2$ versus $load_1$ with $N = 4$ and $\sigma = 0.4$

A similar result can be found in the third case. This is when $\theta_1 < 1$, rewritten as $\frac{\sigma}{\mu_1} > \frac{1-\sigma}{\mu_2}$ or server 2 has a heavier load. The stability condition (31) becomes

$$\gamma < \frac{1-\theta_1+2b}{1-\theta_1-b} \tag{41}$$

or further simplified as

$$\sigma\lambda < \mu_1. \tag{42}$$

In both cases, our stability condition is governed by the queue with the heavier load. This is what is expected, since when the queue with the heaviest load is stable, automatically the other queue with the lower load is also stable. Notice here already that when the load is balanced we get the maximum total load ($\theta_1 = 1$).

We now look at intermediate values of N. Figures 5, 6 and 7 show the influence of the relative frequency distribution. In all three figures we have plotted γ_{sup} versus σ. Here, γ_{sup} is the least upper bound or supremum of the set of γ values where the system is stable and σ represents the relative frequency distribution of the type of the customers. Fig. 5 shows the case where we have a symmetric system ($\mu_1 = \mu_2$) and the types of the customers in the arrival stream are uncorrelated ($K = 1$). In this case, $\sigma = \phi$ and $\sigma = 1-\phi$ ($0 < \phi < 1$) lead to the same results. The key observation to understand the latter is that for the operation of the system the exact types of the N first customers are irrelevant if both servers have the same service rate μ. When the N first customers are all of the same type, the system only processes one customer anyway with the same service rate μ, independent of the type of customer. In fact, a system with $\sigma = 1 - \phi$ can be conceived as a system with $\sigma = \phi$ whereby the names of the types 1 and 2 have been "swapped". Thus, there exists symmetry in the customer types around the value $\sigma = \frac{1}{2}$. The impact of P-gFCFS is the largest when we reach the maximum for γ_{sup} at $\sigma = \frac{1}{2}$. In Fig. 6 and 7 this symmetry is broken since both customers no longer introduce the same average amount of work that enters the system. Fig. 6 shows the case where we have no longer a symmetric system ($\mu_1 = 2\mu_2$) but the types of the customers in the arrival stream are still uncorrelated. Here, we see that the maximum is shifted. This maximum is now at

$$\sigma_{max} = \frac{\mu_1}{\mu_1 + \mu_2} \tag{43}$$

or rewritten

$$\frac{\sigma_{max}}{\mu_1} = \frac{1 - \sigma_{max}}{\mu_2} \tag{44}$$

which in words means that the maximum total load is reached when both customers introduce the same average amount of work that enters the system. In Fig. 7, the asymmetric case ($\mu_1 = 2\mu_2$) where the types of the customers in the arrival stream are correlated, is plotted. Here, we notice that the correlation has indeed no influence on the σ_{max} corresponding with the maximum total load (in both figures 6 and 7, this maximum is at $\sigma = \frac{2}{3}$). But the correlation has an influence on the *value* of the maximum. From figures 5-7 we see that impact of P-gFCFS is most noticeable when the relative load distribution is in balance. If the relative load distribution is totally out of balance the impact of P-gFCFS becomes negligible, which is also intuitively clear since we then approach a system with almost only one type of customers and thus a single server system.

Fig. 8 shows the influence of correlation. In this figure we have plotted γ_{sup} versus parameter K, which gives an indication about the correlation in the system as discussed in Section III-A. Here, we see clearly that the throughput is larger when the types of the customers alternate more in the arrival stream. When $K = \frac{1}{2}$ (when the types arrive constantly alternating), we get the largest total load. Notice that this is not the same value for all different N but the difference is negligible. On the other hand when $K = \infty$, there only arrives one type of customer and we get a single server system and the limit is thus 1. Notice that correlation can have a devastating impact on the total load of the system with P-gFCFS

Fig. 9 and 10 show the influence of the load of one type of customer on the load of the other type of customer. In both figures we have plotted $load_2$ ($= \frac{1-\sigma}{\mu_2}$) versus $load_1$ ($= \frac{\sigma}{\mu_1}$). The $load_2$ in both figures is the least upper bound of the set of $load_2$ values where the system is stable, for a given $load_1$ value. In Fig. 9 we look at the case with uncorrelated types of the customers in the arrival stream ($K = 1$). Here, we see that for $N = 2$, $load_1$ has a huge impact on $load_2$. This impact decreases when N becomes larger. In traffic context is this exactly what we wanted to become with the filter lanes. We wanted to decrease the impact of the vehicles going to destination 1 on vehicles with destination 2 and visa versa. In Fig. 10, the types of the customers in the arrival stream are correlated and we have a parameter $N = 4$ and $\sigma = 0.4$. In this figure we look at the impact of correlation. We can clearly see that in our system, correlation has a devastating impact. It even undoes the impact that or parameter N has on the system.

V. Conclusions and Future Research

In this paper, we have analysed the stability condition of a continuous-time queueing model where two types of customers, both to be served by their own dedicated server, are accommodated in one common FCFS queue with an exception for the first N customers of the system. We have shown the positive impact on the total maximum load of the relaxation of this condition (the parameter N of our P-gFCFS service discipline). We have also deduced that we can achieve the largest maximum total load when the load is balanced between both servers. It is also around this value, our parameter N has the largest impact. When the loads are totally out of balance, our parameter N has almost no impact at all. We have also shown that if there is a lot of correlation between the types of the customers in the arrival stream, this has an devastating impact on our system and can even undo the positive impact of the parameter N (filter lanes). Finally we have also shown that in our system the load of one type of customer can have a big impact on the maximum allowable load of the other type. Here, the parameter N also helps to minimize the impact of one type of customer on the other.

Our future research goal is to go beyond the stability analysis of P-gFCFS. A first interesting extension could be the tail probabilities of the number of customers in the system (or at least some approximation). And especially in traffic context, it would be interesting to be able to set the parameter N so that only a certain percentage of the customers has to wait longer than a certain period of time.

References

[1] T. Van Woensel and N. Vandaele, "Empirical validation of a queueing approach to uninterrupted traffic flows," *4OR, A Quarterly Journal of Operations Research*, vol. 4, pp. 59–72, 2006.

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

[3] 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 *Analytical and Stochastic Modeling Techniques and Applications*, vol. 6751 of *Lecture Notes in Computer Science*, pp. 14–27, Springer Berlin / Heidelberg, 2011.

[4] 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*, vol. 223(1), pp. 123–132, 2012.

[5] L. Kleinrock, *Theory, Volume 1, Queueing Systems*. Wiley-Interscience, 1975.