ON NEW STRATEGY FOR PRIORITISING THE SELECTED FLOW IN QUEUING SYSTEM

Wojciech Burakowski[1], Halina Tarasiuk[1], Ryszard Syski[2]

[1]*Warsaw University of Technology, Poland*
Institute of Telecommunications
00-665-Warsaw, ul.Nowowiejska 15/19
E-mail: wojtek@tele.pw.edu.pl, halina@tele.pw.edu.pl
[2] *University of Maryland, USA*
Department of Mathematics
College Park
Maryland 20742
E-mail: rns@math.umd.edu

Abstract
The paper refers to the queue management algorithm problem for handling submitted to the system traffic flows with different quality of service. The investigated approach, called QMA-R (Queue Management Algorithm with Reservations), assumes queue place reservations made by consecutive customers entering the system and belonging to the selected flow. The number of reservations for given flow depends on its arrival rate and, in this sense, this method differs from the priority based algorithm having predefined priority levels assigned *a priori* for each flow. The behaviour of the system with QMA-R algorithm strongly depends on parameter *a*, denoting the percentage of the preferred flow in the total submitted traffic. The system keeps priority only in the case when *a* is low. For *a* close to 1 the system lost priority properties and serves all customers equally (as with FIFO discipline).
The analysis provided in the paper is limited to the system with two flows with additional assumptions that the input process is Poissonian as well as the service times constitute negative exponential distribution. For such system, the formulas for mean waiting time are derived. Finally, the effectiveness of the proposed algorithm is compared to the systems with and without priorities.

Keywords
Queuing systems, queue management, priorities

1 INTRODUCTION

Introduction different quality of service for particular flows in a packet-oriented network, like IP or ATM based, demands implementation of some queue management algorithms in the nodes. Thanks to them, the network can handle the selected traffic flows better than other ones and, as a consequence, the packets/ATM cells belonging to these flows can be transferred faster via network.

Commonly recommended approach to apply the priorities into a system is to use non-preemptive or preemptive priority scheme. Additionally, in order to satisfy some fairness requirements, the mechanisms like WFQ (*Weighted Fair Queuing*)-based are recently investigated [1], [2]. These mechanisms are common in this sense that the assigned priority for a given flow does not depend on its arrival rate and is fixed *a priori* in a node (e.g. during set-up phase or by management system).

The considered in the paper problem corresponds to the queue management algorithms allowing us to handle traffic flows submitted to a queuing system with different quality. For instance, such a need can occur in the future *diffserv* QoS IP network (e.g. [4], [5]), when computer data traffic corresponding to different applications forms one flow and, despite this, one wants to sent a stream (sub-flow) with better quality then other ones. Especially, the proposed method can be used for better than best effort service. The investigated method, called QMA-R (Queue Management Algorithm with Reservations), belongs to the arrival rate depended class and it assures that the selected flow affects better service without essential degradation of other ones. It is based on the queue place reservation mechanism, which assumes that consecutive customers entering the system and belonging to the selected flow make the reservations. The provided in the paper analysis is limited to the system with two flows. Additionally, we assume Poissonian input and negative exponential service time distributions. For such system, the formulas for mean waiting time are derived. Finally, the effectiveness of the proposed algorithm is compared with the systems with and without priorities.

The organisation of the paper is the following. Section 2 describes the proposed QMA-R algorithm. The analysis of the system is provided in section 3. Section 4 shows exemplary numerical results illustrating the effectiveness of the proposed method by comparing with the results obtained for the systems with and without priorities. Finally, section 5 summarises the paper.

2 SYSTEM DESCRIPTION

The considered system is depicted on Fig.1. This is a single server system with infinite waiting room, which is fed by two independent flows, say #1 and #2. In order to assign a priority for the flow #1, we propose a strategy based on queue place reservation mechanism. The consecutive arriving customers belonging to the flow #1 make these reservations. The service of the customers of the flow #2 is based on the FIFO discipline.

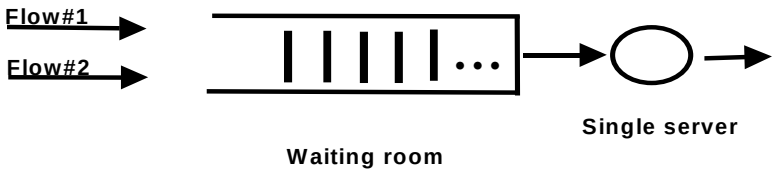


Fig. 1: The studied system

The proposed in this paper method to assign a priority for selected flow is of the arrival rate dependent type and is called QMA-R. It means, that no priority is assigned to this flow *a priori* in the node. The flow affects better service only if some additional conditions corresponding to its arrival rate are performed.

The principles of investigated reservation scheme are explained below and illustrated on Fig. 2. For clarity of the text presentation, let us assume that the consecutive arriving to the system customers of the flow #1 are numbered according to their arrival times; the first arriving customer is no.1, the second is no.2 an so on. The service of the customers from the flow #1 is as follows:

1. Service of the customer no.1 is based on the FIFO rule and is the same as for the customers from the flow #2. However, entering to the system, he reserves the last place in the current queue (just after the place occupied by the customer in question) for the customer no.2. This reserved place is moved up to the top of the queue according to the FIFO.

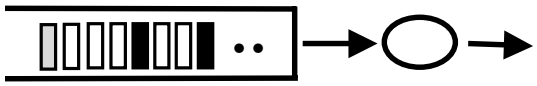
2. When the customer no.2 arrives to the system, he is putting into the reserved for him place (which is not necessary the last occupied place in the current queue). Again, this customer reserves the last place in the current queue for the service of the customer no.3. The reserved place is

kept by the system even in the case when it has reached the top of the queue before arriving of new customer. When such a situation occurs, this new customer is served as customer with assigned the highest priority (non-preemptive).

An example of queue management under investigated reservation scheme is illustrated on Fig.2. Fig.2a shows queue structure just before entering new customer. Notice that in this exemplary queue two places are designated for the customers belonging to the flow #1 but only one of them is currently occupied (the second is just the reservation). The Fig.2b (2c) shows the queue structure just after the customer from the flow#1 (#2) enters the system.

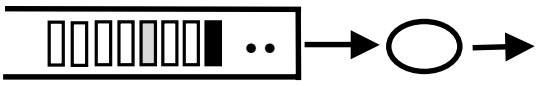
(a) exemplary queue structure just before arriving new customer

(b) queue structure just after arriving customer from the flow#1



The arriving customer from the flow#1 is put into the reserved place in the queue and, simultaneously, the system reserves place (at the end of current queue) for the next customer belonging to this flow

(c) queue structure just after arriving customer from the flow#2. The arriving customer from the flow#2 is put into the end of the queue (FIFO discipline).



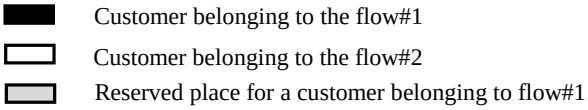


Fig.2: Exemplary queue structure behaviour with place reservation mechanism

3 ANALYSIS OF THE SYSTEM

The objective of the analysis presented in this section is to derive formulas for mean waiting times of the customers belonging to the flow#1 and #2. This analysis has been provided for the system defined in the section 2.1 and under the following additional assumptions:

- arrival process of the customers from the flow #1 (#2) follows the Poissonian distribution with parameter λ_1 (λ_2),
- service time of the customers from the flow#1 (#2) follows negative exponential distribution with parameter $1/\mu_1$ ($1/\mu_2$).

Let us denote:

W_1 (W_2): mean waiting time of the customers belonging to the flow #1 (flow#2) in the studied system (e.g. with the place reservation mechanism);

W_{FIFO}: mean waiting time of the customers in the system with FIFO discipline.

Assuming place reservation mechanism in the system one can expect smaller (higher) waiting times of the customers from the flow#1 (#2) than it happens in the system with the FIFO discipline. Anyway, since these systems are work-conserving [3], one can write the following formula:

$$W_{FIFO} = \frac{\rho_1}{\rho_1 + \rho_2} W_1 + \frac{\rho_2}{\rho_1 + \rho_2} W_2 \qquad (1)$$

The approximate formula for W_{FIFO} is:

$$W_{FIFO} = \frac{\frac{\rho_1(1-\rho_1)}{\mu_1} + \frac{\rho_2(1-\rho_2)}{\mu_2}}{(1-\rho_1)(1-\rho_2)} \qquad (2)$$

,where $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$ and $\rho_1+\rho_2 < 1$. The formula (2) is the exact formula for mean waiting time in the single server system fed by outputs from two M/M/1 systems.

For calculation of W_1, we define additional parameter $\Delta W = W_{FIFO}-W_1$. The method to evaluate ΔW is described below. The value of W_2 is calculated using (1) in straightforward way.

ΔW calculation

As it was stated before, the customers from the flow#1 could affect shorter (comparing to the FIFO discipline) waiting times in the case the reservation mechanism is applied. This is mainly due to the fact that the arriving customer from the flow#1 not necessary is putting to the end of current queue as it happens in the system with FIFO. In the most optimistic scenario, this customer is served before all these customers belonging to the flow#2 which were in the system and arrived after the time the reservation for the customer in question was made. Therefore, the general expression for the ΔW ($= W_{FIFO} - W_1$), has the following form:

$$\Delta W = (1/\mu_2)\sum_{i=1}^{\infty} iP(T=i) \qquad (3)$$

, where

P(T=i) denotes the probability that the random variable T, describing number of the customers from the flow#2 which are served after the customer from the flow#1 but arrived to the system earlier than this customer, takes the value i (i=0,1,2,...).

The values of the P(T=i) (for i=0,1,…) are calculated as follows:

$$P(T=i) = P(N=i)\times P(K \geq i+2) + P(N \geq i)\times P(K=i+1) \quad (4)$$

,where

P(N=n): denotes the probability that random variable N, describing number of customers from the flow#2 who arrives to the system during the inter-arrival time between two consecutive customers from

the flow#1 (this time is done by exponential distribution with the parameter μ_1), takes n (n=0,1,2,..) value.

P(K=k): is the probability that random variable K, describing number of customers being in the system (queue plus server) at the moment a customer from the flow#1 enters the system, takes k (k=0,1,2,...) value.

P(N=n) calculation

Since the customers from the flow#2 arrive to the system according to Poissonian distribution with parameter λ_2, one can write the following equations:

$$P(N=n)=\int_0^\infty P(N=n,t)\lambda_1 e^{-\lambda_1 t}dt=\int_0^\infty \frac{(\lambda_2 t)^n}{n!}e^{-\lambda_2 t}\lambda_1 e^{-\lambda_1 t}dt=\frac{\lambda_1}{\lambda_1+\lambda_2}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^n \tag{5}$$

Notice that, in general case, not all these customers from the flow#2 who arrived to the system before the customer in question, are served later than this customer. This can happen when the reserved place reaches the top of the queue earlier than the customer from the flow #1 enters the system.

P(K=k) calculation

From the point of view of the P(K=k) probabilities, the investigated system with reservation can be considered as belonging to the *M/H$_2$/1* system class. Unfortunately, for such a system there are not exist exact formulas for determining the P(K=k). Therefore, we recommend taking into account two below defined cases, which are:

Case no.1: $\mu_1 \neq \mu_2$, $\lambda_1 \neq \lambda_2$ and $\rho_1=\frac{\lambda_1}{\mu_1}$, $\rho_2=\frac{\lambda_2}{\mu_2}$.

For this case, the approximate formula is the following:

$$P(K=k)=(\rho_1+\rho_2)^n(1-\rho_1-\rho_2) \tag{6}$$

Case no.2: $\mu_1 = \mu_2 = \mu$, $\lambda = \lambda_1 + \lambda_2$.

In this case the system is the *M/M/1* with arrival rate = λ, and mean service time = $1/\mu$. The exact formula for the *P(K=k)* has the following form:

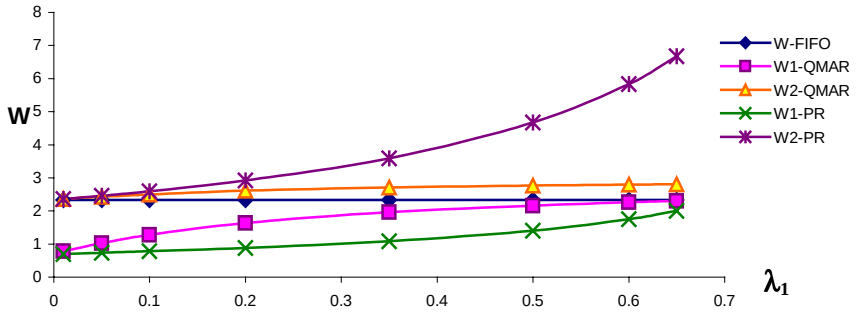
$$P\left(K = k\right) = \left(\lambda / \mu\right)^k \left(1 - \lambda / \mu\right) \quad (7).$$

4 NUMERICAL RESULTS

This section presents exemplary numerical results illustrating some benefits of the proposed QMA-R algorithm (system no.1) comparing to the system with FIFO (system no.2) and the system with non-preemptive priorities (system no.3). The studied system case consists of single server and infinite waiting room (see Fig.1). The system performances will be examined in terms of the mean waiting times corresponding to the flow #1 and #2. The flow #1 is served in the system no.1 assuming reservation mechanism and with assigned higher non-peemptive priority in the systems no.3. Therefore, in these two systems one can expect smaller waiting times for the flow#1 than for the flow#2. In the system no.2 the flow#1 and #2 are served without any priorities, so the corresponding mean waiting times are the same.

Fig 3 shows the mean waiting times characteristics as a function of arrival rate of customers belonging to the flow#1, λ_1, collected for the three above defined systems. These results were obtained assuming that overall load in each of these systems was 0.7. They say that the mean waiting time values of the customers belonging to the flow#1 in the system no.1 comparing to these obtained for the system no.2 and no.3 strongly depend on the percentage of the flow#1 in the total submitted flow. When λ_1 is relatively small one can expect similar service quality for the flow#1 to this offered by the system with priorities (no.3). On the contrary, when λ_1 is rather high the system with reservation mechanism (no.3) behaves similar to the system with FIFO (no.2). Therefore, in the latest case we do not observe non-desirable effect, occurring in the system with priorities (no.3), leading to fast service quality degradation of the customers from the flow #2.

Mean waiting times



$W_{1\text{-}QMAR}$, $W_{2\text{-}QMAR}$ – mean waiting time of customers belonging to the flow#1 (#2) for the system no.1 (with reservation mechanism for the flow#1)
W_{FIFO} – mean waiting time of customers belonging to the flow#1 and #2 for the system no.2 (with FIFO queue discipline)
$W_{1\text{-}PR}$, $W_{2\text{-}PR}$ – mean waiting time of customers belonging to the flow#1 (#2) for the system no.3 (with non-preemptive priorities - flow#1 has priority over flow#2)

Fig.3: Mean waiting times as a function of λ_1, for the system with $\mu_1=\mu_2=1$, $\lambda_1+\lambda_2=0.7$, $\rho=0.7$

Mean waiting times

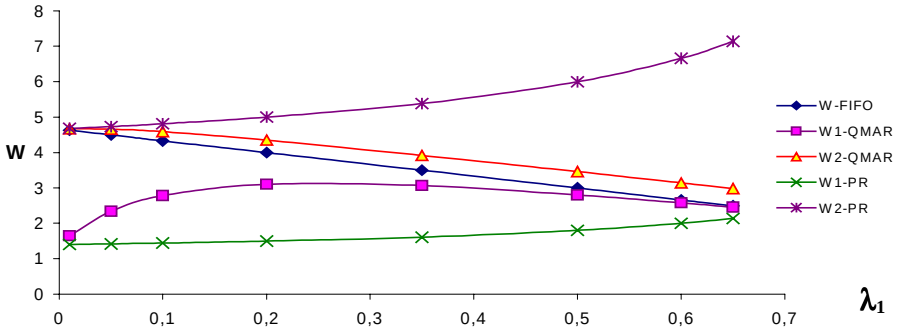


Fig.4: Mean waiting times as a function of λ_1, for the system with $\mu_1=1$, $\mu_2=0.5$, $\rho=\rho_1+\rho_2=0.7$

The mean waiting time characteristics as a function of arrival rate of customers belonging to the flow#1, λ_1, assuming different values of parameters μ_1 (=1) and μ_2 (=0.5) are depicted in the Fig.4 and 5. These results are similar to these from the Fig.3 and the conclusions are the same.

Mean waiting times

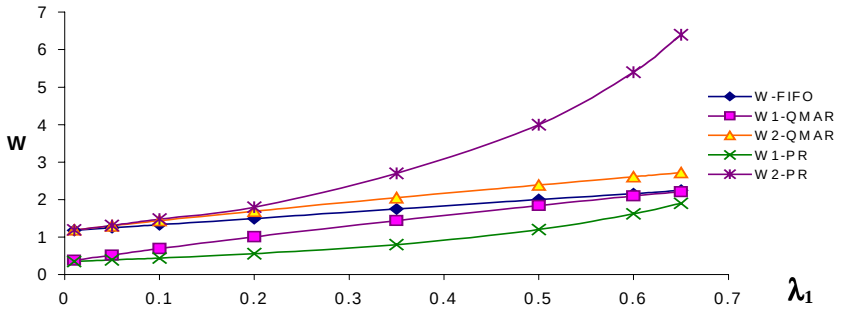


Fig.5: Mean waiting times as a function of λ_1, for the system with $\mu_1=1$, $\mu_2=2$, $\rho=\rho_1+\rho_2=0.7$

Network scenario

For checking the performances of the recommended algorithm in a network environment we show some results corresponding to the network configuration, depicted on Fig.6. In this network scenario, two test flows, say #T0 and #T1, are transferred by n servers organised in a tandem. Additionally, to each of these servers is submitted a transient flow. Similarly to the studies provided for the single server case, at present we also distinguish three cases:

case no.1: the flow#T0 is served with reservation mechanism, the flow #T1 and the transient flows are served without reservation mechanism,

case no.2: all flows in the network are served without priorities (the FIFO discipline is assumed in each server,

case no.3: the flow#T0 has assigned the highest priority, the flow#T1 and the transient flows are served with lower priority.

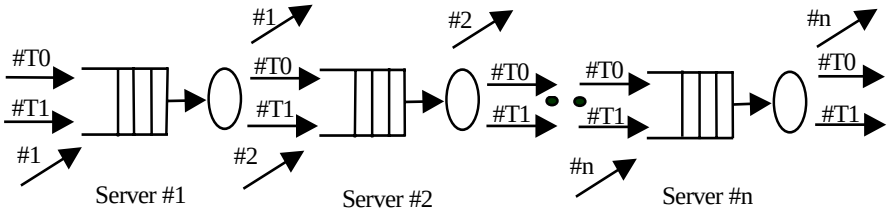
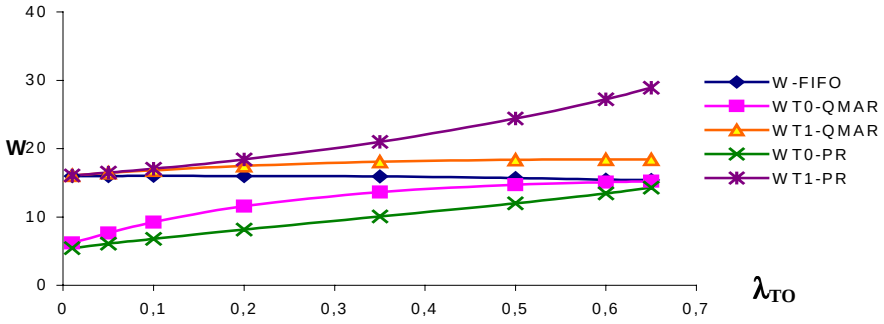


Fig.6: The studied network scenario

The gathered numerical results were obtained using simulation. The cumulative mean waiting time characteristics for the above defined scenario as a function of the arrival rate of the flow#T0, $\lambda_{TO,}$ and corresponding to the network with n=6 servers are depicted on Fig.7. In general, they confirm the conclusions for the single server case. The differences between the considered mean waiting times are now greater and this is caused by the fact that each server degrades the flows in similar way.

Cumulative mean waiting times



$W_{T0\text{-}QMAR}$, $W_{T1\text{-}QMAR}$ – mean waiting time of customers belonging to the flow#T0 (#T1) for the case no.1 (with reservation mechanism for the flow#T0)
W_{FIFO} – mean waiting time of customers belonging to the flow#T0 and #T1 for the case no.2 (with FIFO queue discipline)
$W_{T0\text{-}PR}$, $W_{T1\text{-}PR}$ – mean waiting time of customers belonging to the flow#T0 (#T1) for the case no.3 (with non-preemptive priorities - flow#T0 has priority over flow#T1)

Fig.7: Mean waiting times as a function of λ_{T0}, for the system with $\mu_{T0}=\mu_{T1}=\mu_i=1$, $\lambda_{T0}+\lambda_{T1}+\lambda_i=0.7$, $\lambda_{T1}=\lambda_i$ (i=1,...,n), ρ= 0.7, n=6

5 CONCLUSIONS

In the paper a new strategy, called QMA-R, based on queue place reservation mechanism for prioritising selected flow was proposed. The formulas for mean waiting time in such a system were derived assuming Poissonian input and exponential service time distribution. The included exemplary numerical results showed some benefits of the proposed mechanisms comparing to these for the system without priorities and with non-preemptive priorities. Applying this strategy to a flow one can expect that when the traffic produced by this flow is a small percentage of the total load in the system then this flow affects privileged service quality similar to this observed for the flow with the highest priority in the system. On the contrary, then this percentage is rather high then the system behaves as the system without priorities.

The method can be extended to the case with more than two types of flows, simply by introduction different queue threshold values for each flow.

References

[1] S. Keshav, An engineering approach to computer networking, chapter 9: Scheduling, Addison-Wesley, 1997

[2] Final report COST 242, Broadband network teletraffic: Performance evaluation and design of broadband multiservice networks (J. Roberts, U. Mocci, J. Virtamo eds.), Lectures Notes in Computer Science 1155, Springer 1996.

[3] L. Kleinrock, Queuing Systems – Applications, Addison Wesley, 1976.

[4] S. Blake et al., An Architecture for Differentiated Services, Internet RFC 2475, December 1998.

[5] Y. Bernet et al., Differentiated Services, Internet Draft, draft-ietf-diffserv-framework-0.2.txt, February 1999.