Generalization of preemptive and non-preemptive priority queues

Tom Maertens, Joris Walraevens and Herwig Bruneel

Ghent University – UGent
Department of Telecommunications and Information Processing
SMACS Research Group
Address: Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Phone: +32-9-2648901, Fax: +32-9-2644295
E-mail: {tmaerten,jw,hb}@telin.UGent.be

Priority scheduling is still a hot topic in queueing theory. In many queueing systems, real-life situations involving human beings as well as computer systems, different types of customers require different kinds of services. In telecommunication systems, for example, one can think of packets that belong to a video conference application. To guarantee the quality of such conference, it is necessary that these packets are not being held up too much. When all packets are transmitted from a certain network node in the order in which they arrive at this node, no *service* or *delay differentiation* is possible between different types of packets. By prioritising the transmission of packets that belong to the video conference (in some or even in all the nodes), it is possible to achieve the required service differentiation.

Priority scheduling disciplines can be divided into two categories with respect to how they deal with arriving customers having a higher priority than the customers that are currently being served. Particularly, in a *non-preemptive* priority queue, the service of a lower-priority customer is not interrupted when a higher-priority customer arrives at the system. Once the service of the lower-priority customer is finished, the server starts servicing the higher-priority customer. In a *preemptive* priority queue, on the other hand, the service of a lower-priority customer will be interrupted at once if a high-priority customer arrives, and will not be resumed until the system is again void of higher-priority customers. It is easily seen that the preemptive category is favourable to higher-priority customers, because they are not influenced by lower-priority customers at all. On the other hand, with non-preemptive priority, low-priority customers are at least sure of being served completely once their service is started. Much research has been done on priority scheduling disciplines, non-preemptive as well as preemptive (see e.g., [3, 4]). Both categories, however, have several drawbacks in practical applications. Under the non-preemptive category, higher-priority customers may have to wait even when the service of a lower-priority customer has just started, while under the preemptive disciplines, the almost completed service of a lower-priority customer may be interrupted due to the arrival of higher-priority customers (pos-

sibly causing a large extra delay).

In our research, we propose a priority scheduling discipline in which the two above-mentioned situations are avoided as much as possible. In particular, we introduce a parameter γ which is defined as the fraction of the service time that already has to be elapsed in order that the service of a lower-priority customer is no longer interrupted when a higher-priority customer arrives at the system. In other words, the ratio of the elapsed service time of a lower-priority customer upon arrival of a higher-priority customer to its total service time is compared with γ. If that ratio is smaller than γ, the service of the lower-priority customer is interrupted; otherwise, its service is completed before the service of a higher-priority customer can start. Note that the preemptive and non-preemptive priority disciplines are two special (extreme) cases of our newly defined discipline, namely they correspond with $\gamma = 1$ and $\gamma = 0$ resp.

This new priority scheduling discipline resembles the discretionary priority discipline introduced in [1] and studied in detail in [2]. In the latter priority discipline, however, the authors put an absolute threshold on the elapsed service time of lower-priority customers to interrupt their service, while we propose a relative threshold, which makes more sense.

We have done a preliminary simulation study of a discrete-time queue with the newly proposed discipline. We have assumed two priority classes, and a jointly binomial distribution for the number of high- and low-priority arrivals in a slot. The performance measure is a weighted linear cost function of the average delays of both priority classes. For deterministic service times, we have observed a cost percentage gain of up to 8% as opposed to preemptive and non-preemptive priority. For variable service times (geometric distribution), we have observed even larger cost gains.

We conclude that the new priority discipline seems promising. However, an analytic study of the performance of this scheduling is necessary in order to make definite conclusions. We will furthermore have to come up with an alternative when the total service time of the low-priority customers is not known beforehand, as our discipline assumes knowledge of this total service time.

Ackowledgment This work is partly based on the master thesis of Nikolaas Van Heucke.

References

[1] B. Avi-Itzhak, I. Brosh, and P. Naor. On discretionary priority queueing. *Zeitschrift für Angewandte Mathematik und Mechanikz*, 44(6):235–242, 1964.

[2] K. Kim and K.C. Chae. Discrete-time queues with discretionary priorities. *European Journal of Operational Research*, 200(2):473–485, 2010.

[3] J. Walraevens, B. Steyaert, and H. Bruneel. Delay characteristics in discrete-time gi-g-1 queues with non-preemptive priority queueing discipline. *Performance Evaluation*, 186(1):182–201, 2002.

[4] J. Walraevens, B. Steyaert, and H. Bruneel. Analysis of a discrete-time preemptive resume priority buffer. *European Journal of Operational Research*, 186(1):182–201, 2008.