A discrete-time queueing model with constant service times and blocking

H.Bruneel, W. Mélange, D. Claeys, J. Walraevens
Ghent University, Departement of Telecommunications and Information Processing
E-mail : hb,wmelange,dclaeys,jw@telin.Ugent.be

We analyze a discrete-time queueing model where two types of customers, each having their own dedicated server, are accommodated in one single FCFS queue. Service times are deterministically equal to $s \geqslant 1$ time slots each. Customers are served in their order of arrival, regardless of the class they belong to. We call this service discipline "global FCFS". New customers enter the system according to a general independent arrival process, but types of consecutive customers may be non-independent in a first-order Markovian way. Specifically, we denote by α the probability that the next customer has *the same type as the previous one*, and by $1-\alpha$ the probability that the next customer belongs to *the opposite type as the previous one*. The parameter α can then be considered as a measure of the degree of class clustering in the arrival process, and is referred to as the "cluster parameter".

The motivation for this work are some every day situations where this kind of blocking occurs. A first instance are traffic jams. One reason for traffic jams are traffic junctions. Consider for instance the following situation : vehicles approach a junction with two possible destinations (1 and 2). 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 of the destination. Similarly, in switching nodes of telecommunication networks, information packets with a given destination of node 1 may have to wait for the transmission of packets destined to node 2 that arrived earlier, even when the link to node 1 is free, if the arriving packets are accommodated in so-called input queues according to the source from which they originate (the well-known HOL-blocking effect).

We use the generating function method to calculate the main performance measures of the system, which are the mean system content $E[u]$, the mean system delay $E[d]$ and the mean unfinished work $E[w]$. We derive the probability generating functions of the system content and unfinished work. By applying (the discrete-time version of) Little's law, the mean system delay of a customer can be obtained.

The results confirm the strong impact of "class clustering" in the arrival stream on the stability and the main performance measures of the system. For example, the form of the stability condition is given by

$$\lambda s < 2 - \alpha$$

where λ is the mean (per slot) arrival rate. This form reveals that the maximum achievable throughput of the system, expressed in work units per slot, is very directly determined by the degree of class clustering in the arrival process as described by the cluster parameter α. Specifically, when α varies from 0 to 1, the maximum throughput is divided by a factor 2.

The results also illustrate the effects of the lengths of the service times and the burstiness in the customer arrival process. This model has some interesting scaling properties.