Motivated by various applications in logistics, road traffic and production management, we investigate two versions of a tandem queueing model in which the service rate of the first queue can be controlled. The objective is to keep the mean number of jobs in the second queue as low as possible, without compromising the total system delay (i.e. avoiding starvation of the second queue). The balance between these objectives is governed by a linear cost function of the queue lengths. Such models can be formulated as a Markov Decision Process for which the optimal decision in each state can be numerically computed.

In the first model, the server in the first queue can be either switched on or off, depending on the queue lengths of both queues. This model has been studied extensively in the literature. Obtaining the optimal control is known to be computationally intensive and time consuming. In previous research it has been shown that a switching strategy is optimal. The structure of the switching strategy can be divided into two cases, which depend on the difference in service rate between both servers. We are particularly interested in the scenario that the first queue can operate at larger service rate than the second queue. This scenario has received less attention in literature. Numerical results indicate that the optimal switching curve is rather flat. We therefore investigate the efficacy of fixed-threshold strategies and formulate the system as a controlled fluid problem. In this setting, finding the optimal threshold value is much less computationally demanding than solving the original MDP problem.

Next, we investigate the case in which the first queue is a (controllable) batch server. This means that the first server transfers jobs in batches rather than individually. In this system we assume that the service rate is independent of the batch size. We show that this problem is similar to the initial tandem queueing model with controllable service rate at the first stage. The optimal strategy of the batch model can be characterised as a switching strategy. Moreover, the switching line determines the optimal division of jobs over the two queues. Specifically, the optimal batch size can be characterised as a "jump to" policy. The switching line represents the optimal state; each time a transition occurs, the size of the batch will be such that the process "jumps to" this line. We will prove that this "jump to" strategy is optimal. This allows us to approximate the batch model in terms of the above fluid control problem as well. We compare results of the optimal strategy for both implementations and show that the fluid control formulation is a good approximation for the MDP formulation.

We illustrate the appropriateness of our approximations using simulations of both models.