Queueing Syst (2017) 87:269–291
https://doi.org/10.1007/s11134-017-9554-x

# Optimal dispatching in a tandem queue

**D. van Leeuwen**[1] · **R. Núñez Queija**[1,2]

Received: 14 October 2016 / Revised: 22 September 2017 / Published online: 27 October 2017


**Abstract** We investigate a Markovian tandem queueing model in which service to the first queue is provided in batches. The main goal is to choose the batch sizes so as to minimize a linear cost function of the mean queue lengths. This model can be formulated as a Markov Decision Process (MDP) for which the optimal strategy has nice structural properties. In principle we can numerically compute the optimal decision in each state, but doing so can be computationally very demanding. A previously obtained approximation is computationally efficient for low and moderate loads, but for high loads also suffers from long computation times. In this paper, we exploit the structure of the optimal strategy and develop heuristic policies motivated by the analysis of a related controlled fluid problem. The fluid approach provides excellent approximations, and thus understanding, of the optimal MDP policy. The computational effort to determine the heuristic policies is much lower and, more importantly, hardly affected by the system load. The heuristic approximations can be extended to models with general service distributions, for which we numerically illustrate the accuracy.

**Keywords** Tandem queues · Optimal control · Fluid approximation · Batch service

**Mathematics Subject Classification** 60J20 · 90C40

✉ D. van Leeuwen
D.van.Leeuwen@cwi.nl

R. Núñez Queija
nunezqueija@uva.nl

1 Centrum Wiskunde en Informatica, Amsterdam, The Netherlands

2 Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

## 1 Introduction

We investigate a controllable two-stage tandem queue: the first stage represents a storage buffer where jobs can be kept before being transferred to the second stage. Our main motivation for this model comes from road traffic control, where one can avoid accumulation of traffic by reducing the upstream traffic flow [15]. It is assumed that the buffer is large enough so that it is reasonable to model it with infinite storage capacity. The second stage represents the service bottleneck for which we want to maintain a small number of waiting jobs. In practice, service at the first station may not be limited to one job at a time. For example, in manufacturing and production planning, several items may be produced or delivered at the same time. In road networks, cars often drive in platoons from one road segment to the next. In many applications, it is reasonable to assume that the service rate is independent of the number of jobs that are jointly processed. (Think, for example, of platoons of vehicles jointly driving from a buffer segment to the critical segment.) We seek an optimal trade-off between a reduction in the number of jobs at the second stage and the additional delay caused by keeping jobs in the first stage. The optimal point of operation is determined by minimization of a cost function that accounts for waiting time in the buffering stage as well as waiting at the critical stage. Arrivals to this system are modelled by a Poisson process, and service times in both queues are exponentially distributed, which facilitates a formulation as a Markov Decision Problem (MDP). Solving the MDP to optimality is in general computationally prohibitively demanding. Our main objective in the paper is therefore to first identify the main structure of the optimal strategy and then develop two heuristic approaches that closely approximate the optimum. Our heuristics will be based on the analysis of a related controlled fluid model and provides intuition for the optimal decision structure.

The proposed model is rather well understood for the single-service model, in which the first server either serves a single job or idles. This setting has been considered in [1,5,13,17] and is the basis for our analysis of the batch service model; we will discuss these references in more detail in Sect. 2.

Several papers have investigated control of similar tandem models, of which we discuss the most relevant. In [16] an inventory control system has been analysed for various control policies in which both the first and the second station can be controlled. The fact that in that paper the first station represents an inventory level, which can be negative, fundamentally changes the analysis. It is worth noting that with an appropriate translation, their special order-to-stock policy is mathematically equivalent to our single-service model with a fixed threshold at the second queue. In [18], a fuzzy control mechanism is used that computes the decision at each state based on expected reward versus holding costs. This approach is of similar spirit to our fluid-based approximations for which we also use expected costs to approximations, the threshold value.

Batch service models with control for single queues have been studied, for example, in [10,11]. The optimal batch size is determined by a trade-off between costs of a service initiation, and the waiting time costs of jobs in the system. In the present paper, we do not consider costs for service initiation, but costs are related to lost capacity. Capacity is lost when the second server becomes idle, while the first queue is not

empty. Alternatively, the model in [7] charges costs for abandonment due to impatient customers in a batch service queue. In our model, the adverse effect of being delayed in the first queue is indirectly penalized by the fact that the second queue may idle unnecessarily.

Most of our attention will be targeted at the batch service model. The approximations we propose have natural counterparts for the single-service tandem model as well. To avoid overly repeating discussions, we will not derive these in detail, but on several occasions we will briefly refer to the similarities and differences of the two models, and we will also use the single-service model to illustrate the applicability of our heuristics for non-exponential service times.

The remainder of the paper is organized as follows. In Sect. 2 we first provide the necessary background summarizing previous work. We then describe the batch tandem control model in Sect. 3 and cast this problem into an MDP framework, investigating its structural properties. To gain more intuition for the decision structure of the optimal policy, we proceed with a fluid formulation in Sect. 4. Section 5 describes a method to approximate the optimal control strategy of the MDP formulation using the fluid model. We illustrate the accuracy of the approximations in Sect. 6. Subsequently, we generalize the heuristic approximations for use in models with more general (phase-type) service distributions in Sect. 7. The final section contains conclusions and ideas for further investigation.

## 2 Previous work

We have previously investigated the tandem queueing model with single services in the first queue [15]. In that case, control reduces to deciding whether to switch the first server on or off, depending on the queue lengths of both queues. For exponential service time distributions, the optimal strategy is prescribed by a switching curve [5,13]. We now discuss the structural properties of this single-service model, so as to later compare with the batch service model. The switching curve dictates a (dynamic) threshold on the number of jobs in the second queue that depends on the number of jobs in the first queue. If the number of jobs in the second queue is below the switching curve, the policy prescribes transferring new jobs from the first queue to the second queue. The shape of the switching strategy shows a sharp dichotomy [1,15], depending on which queue has the highest service rate. Most theoretic results concentrate on the case where the first queue is the bottleneck (has lower service rate than the second queue). Using asymptotic analysis, Avram [1] shows that the optimal strategy has a linearly increasing switching curve (in the fluid limit).

The more relevant regime for us, in which the first queue can operate at a larger service rate than the second queue, has received less attention in literature. In that case the optimal switching curve is rather flat [15] and thus requires a different approach than that in Avram [1]. (This will be illustrated later in Fig. 2.) The flat switching curve suggests that the optimal policy may be approximated by a fixed-threshold policy [15]. Contrary to the results in [14] for a two-station tandem queue *with admission control,* we found that performance is quite sensitive to the exact threshold value in our case. Computing the best threshold policy using matrix-geometric analysis is much less

computationally demanding than solving the original MDP problem [15]. We showed that the best fixed-threshold policy can approximate the flat optimal switching curve rather well, and is able to achieve comparable performance to the optimal MDP solution. Unfortunately, for heavy loads, the matrix-geometric approach also suffers from long computation times.

In [15] we also extended the fixed-threshold approximation to the batch service model of this paper: the server at the first queue transfers jobs in batches rather than individually, the service rate being independent of the batch size. The optimal strategy in this problem has a similar structure as the single-service tandem queueing model with a rather flat switching strategy, which is natural in view of the augmented service capacity in the first queue due to batch services. Since we do not impose any limits on the batch sizes, the optimal strategy is a "jump-to" policy, in which the optimal batch size is such that when the batch is transferred from the first to the second queue, the state is exactly on the switching curve.

In this paper we continue our investigation of the batch service model. To overcome the computational burden of the matrix-geometric approximation for large loads, we develop new approximations using a fluid analysis motivated by [1] but with an alternative scaling in which the first queue may contain a large number of jobs, while the 'critical' second queue remains of moderate size. The randomness in the second queue determines the fluid dynamics of the first stage. In Sect. 4 we will discuss differences between our fluid approximation and existing fluid limits. The fluid-based approach results in two heuristic strategies that provide excellent approximations for a broad range of parameter values, while the computation time is quite insensitive to the system load.

## 3 Model description and preliminary analysis

In our tandem model, we control the service at the first station, which is operated by a server providing service in batches. The control mechanism allows us to choose the batch size at transfer time instants. For analysis purposes we assume that jobs arrive to the first queue according to a Poisson process at rate $\lambda$ and jobs are processed in batches with rate $\mu_1$. (We will extend our approximations to include general service time distributions.) The rate of service is independent of the batch sizes, and the sizes of the batches can be chosen arbitrarily up to the number of jobs in the first queue. After service in the first queue, jobs proceed to the second queue, for which the service rate is denoted by $\mu_2$, handling jobs one by one. A graphical representation of the batch service tandem model is shown in Fig. 1. The salient feature of the model is that the first queue processes $a \in \{0, 1, \ldots, K\}$ jobs at the same time instead of handling jobs one by one. The main goal is to *dynamically* determine (or approximate) the optimal batch size based on the state of the system.

So as to formulate our tandem model in terms of an optimization problem, we introduce a cost function for jobs in the system. We have two types of costs, (1) waiting costs $c_{\text{wait}}$, which are incurred *per job in the system* and *per unit of time,* and (2) location costs for jobs queueing at the second station, $c_{\text{loc}}$, which represent the costs of residing in the service area of station 2. Thus, the waiting cost at the first server is

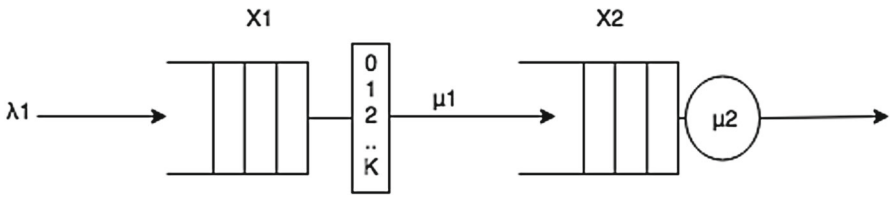


**Fig. 1** Graphical representation of the tandem queue with batch service in the first queue

$c_1 = c_{\text{wait}}$ per job per unit of time, and at the second station it is $c_2 = c_{\text{wait}} + c_{\text{loc}}$. Naturally, we assume $0 < c_1 < c_2$. Due to larger costs at station 2, it is advantageous to hold customers in queue 1 rather than in queue 2. However, one should avoid the situation where station 2 empties while station 1 still has a backlog: the resources at station 2 would be wasted in that case. We seek an efficient trade-off between these two effects.

We first formulate our optimization problem as an MDP and then investigate the structural properties of the optimal policy using results from the literature and numerical experiments. In our experiments, we numerically solved the MDP, which in many cases took days of computation time, so we have the ideal reference for comparison with our fluid approximations.

### 3.1 MDP formulation

Our discrete-time Markov Decision Process consists of the quadruple $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{C}\}$, where $\mathcal{S}$ represents the state space of the system and is defined as $x = (x_1, x_2) \in \mathbb{N}_0^2$, with $x_k$ as the number of jobs at station $k$, $k = 1, 2$. Each state $x$ has a set of allowed actions, or batch sizes, $a \in \mathcal{A}_x = \{0, \ldots, x_1\}$, where $\mathcal{A}_x \subset \mathcal{A}$ and $x \in \mathcal{S}$. The function $p^a(x, y) \in \mathcal{P}$ defines the transition probability from state $x$ to state $y$ for action $a$, where $x, y \in \mathcal{S}$ and $a \in \mathcal{A}_x$. The cost function $c^a(x) \in \mathcal{C}$ states the costs for action $a$ in state $x$.

An optimal strategy satisfies Bellman's equation [2,9]:

$$V^*(x) + g^* = \min_{a \in \mathcal{A}_x} \left\{ c^a(x) + \sum_{y \in \mathcal{S}} p^a(x, y) V^*(y) \right\} \text{ for } x \in \mathcal{S}, \tag{1}$$

where $g^*$ and $V^*(x)$ give the optimal average reward and value function. The decision rule can be determined by

$$f(x) \in \operatorname{argmin}_{a \in \mathcal{A}_x} \left\{ c^a(x) + \sum_{y \in \mathcal{S}} p^a(x, y) V^*(y) \right\} \text{ for } x \in \mathcal{S}, \tag{2}$$

where $V^*(x)$ satisfies $V^*(x) + g^* = c^f(x) + \sum_{y \in \mathcal{S}} p^f(x, y) V^*(y)$. Note the slight abuse in notation in writing $c^f(x)$ and $p^f(x, y)$ instead of $c^{f(x)}(x)$ and $p^{f(x)}(x, y)$. Our goal is to minimize the average cost.

To determine the optimal strategy in our tandem queue we use Eq. (2), where $c^a(x)$, with $x = (x_1, x_2)$, is given by $c_1x_1 + c_2x_2$. We use uniformization to discretize the Markov chain as described in Lippman [8]. The transition probabilities $p^a(x, y)$ are determined by the transition rates in each state.

Letting $\lambda + \mu_1 + \mu_2 = 1$ without loss of generality, for $x = (0, 0)$ we have

$$p^a(x, y) = \begin{cases} \lambda & \text{if } y = (1, 0) \\ \mu_1 + \mu_2 & \text{if } y = (0, 0) \end{cases},$$

and, for $x = (x_1, x_2) \neq (0, 0)$,

$$p^a(x, y) = \begin{cases} \lambda & \text{if } y = (x_1 + 1, x_2) \\ \mu_1 & \text{if } y = (x_1 - a, x_2 + a) \\ \mu_2 & \text{if } y = (x_1, x_2 - 1) \end{cases} \quad \begin{array}{l} \text{for } a \in \{0, \ldots, x_1\}. \\ \text{or } x = y = (x_1, 0) \end{array} \tag{3}$$

For this system, there is always a strategy that yields a stable Markov chain as long as $\lambda < \mu_2$, irrespective of the value of $\mu_1 > 0$. This is obvious, as we can choose to always serve all jobs present in queue 1 in a single batch, no matter how many there are, thereby emptying the first queue at times dictated by an independent Poisson process with rate $\mu_1$. As long as $\lambda < \mu_2$, the second server will be able to stabilize the system in the long run.

We use Successive Approximation (SA) to find the policy that minimizes average costs for each state:

$$\begin{aligned} V_n^*(x) = c_1x_1 + c_2x_2 + \lambda V_{n-1}(x_1 + 1, x_2) + \mu_2 V_{n-1}(x_1, (x_2 - 1)^+) \\ + \mu_1 \min_{a \in \mathcal{A}_x} \left\{ V_{n-1}^( x_1 - a, x_2 + a) \right\}, \end{aligned}$$

and

$$f_n^*(x) \in \operatorname{argmin}_{a \in \mathcal{A}_x} \left\{ c(x) + \sum_{y \in \mathcal{S}} p^a(x, y) V_{n-1}^*(y) \right\}.$$

We initiate the recursion with $V_0^*(x) = 0$ for all $x \in \mathcal{S}$ and $n \in \mathbb{N}$.

### 3.2 Structural properties

In this section, we investigate structural properties of the optimal policy for our batch model. For reference, we compare these to the results of the model with individual services. The numerical results show a similar switching strategy. More specifically, we observe a 'jump-to' strategy which we discuss in detail.

As explained in Sect. 2, the optimal strategy of the tandem control model with single services at the first stage is characterized by a switching strategy [13,15], which divides the state space into two areas separated by the switching curve, as illustrated below in Fig. 2. It is optimal to block service at the first station above the curve, and below

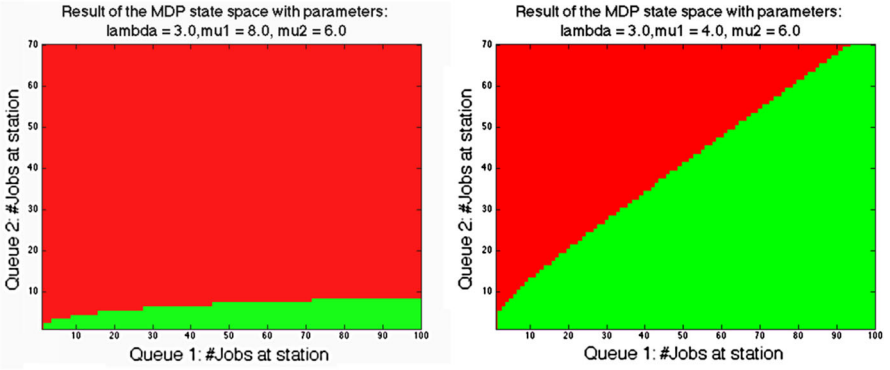


**Fig. 2** Optimal decision strategy at each state for the single-service model distinguishing the cases $\mu_1 < \mu_2$ and $\mu_1 > \mu_2$. The red colour represents blocking and the green colour corresponds to service in the first queue (Color figure online)

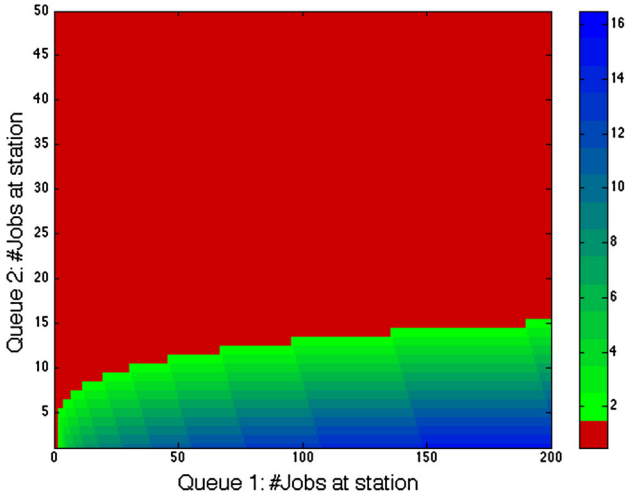


**Fig. 3** Optimal decision strategy at each state for the batch service model with parameters $\lambda = 4$, $\mu_1 = 2$ and $\mu_2 = 6$. Each colour represents a different optimal batch size for the current state (Color figure online)

the curve it is optimal to continue service. We see a similar switching strategy for the batch service model in Fig. 3. In that case, above the curve it is optimal not to transfer any jobs from the first to the second queue and below the curve it is optimal to transfer a batch of jobs. The optimal size of the batch can also be determined from the switching curve, as depicted in Fig. 4: the optimal batch size is determined by the aggregate number of jobs in the two queues. If the total number of jobs is $x_1 + x_2 = N$, the optimal action is to serve $a$ jobs in the first queue such that $(x_1 - a, x_2 + a)$, which also has $N$ jobs in total, is on the switching curve. Should this value of $a$ be negative (this happens when $(x_1, x_2)$ is in the red area), then no jobs should be served in the first queue. With the optimal strategy, the process thus jumps to the switching curve after each completion of a batch.

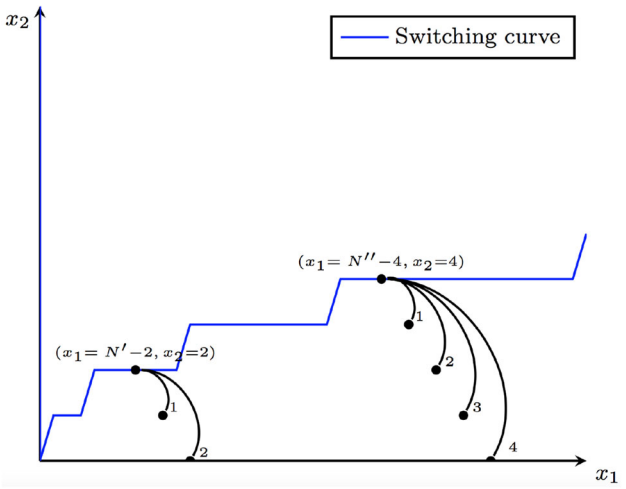


**Fig. 4** Graphical representation of the optimal 'jump-to' structure in the batch tandem queue for fixed $k$

The single-service model shows a clear distinction between the cases $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$. For $\mu_1 < \mu_2$, the curve is rather flat, while it has a distinctive linear shape when $\mu_1 > \mu_2$. For the batch model, there is no such discrepancy: the switching patterns are rather flat for any choice of $\mu_1$, as illustrated in Fig. 3. This is not surprising in view of the (unlimited) increase of service capacity in the first queue due to the batch services, no matter how small $\mu_1$ is.

## 4 Fluid model description

We will reformulate our queueing model as a fluid control problem so as to approximate the optimal strategy found by solving the MDP (using successive approximation). We apply fluid scaling to the first queue, while preserving the queueing behaviour at the second queue. This approach is motivated by the earlier observation that the optimal switching curve is rather flat. Our scaling is different from the standard fluid scaling as first proposed in Kurtz [6]. Since the second queue is not scaled, it maintains its stochastic nature. This randomness is of a different nature than that described in [4] for a model with two queues, where the trajectories of the fluid-scaled components are random. Our scaling is also different from those in the batch service model of [3]. Their first scaling is the standard one of [6] and in the second the batch sizes are scaled, so that the limiting fluid model has jumps.

Our scaling is closest to that described by Robert [12, Ch.9.6]. In that work, however, the unscaled components have stationary distributions that do not depend on the position of the scaled components. In our case, the conditional distribution of queue 2 (the unscaled component) depends on the position of the fluid-scaled size of queue 1. In this paper we do not formally prove the convergence of the scaled stochastic process to the proposed fluid model, as was done in [12]. Instead, we propose the approximation

by investigating local dynamics and illustrate the appropriateness through numerical experiments.

Let us briefly recall the fluid limits in Avram [1] for the *single-service* controlled tandem model. It turns out that for the case $\mu_1 < \mu_2$ the optimal strategy in the fluid model is determined by a linearly increasing switching line, but for $\mu_1 > \mu_2$, the switching line lies on the horizontal axis. This can be understood from the flat, unscaled, switching curve: in the fluid scaling it is indistinguishable from the $x$-axis. We therefore need a different scaling if $\mu_1 > \mu_2$, and the same is true for the batch service model: For the first queue we can apply the usual fluid scaling, but the second queue should remain unscaled.

Formally, the fluid limit for the batch service model is obtained as the limit of a sequence of processes

$$\left\{\left(X_1^{(n)}(t), X_2^{(n)}(t)\right),\ t \geq 0\right\}_{n\geq 1}$$

indexed by $n$, which we take to be integer, as $n \to \infty$. The sequence is determined by the queue length processes of the first and second queue, $X_1(t)$ and $X_2(t)$, respectively. Motivated by the observation from [1] discussed above and justified by numerical experiments that show that the optimal control policy indeed employs an asymptotically flat switching curve, we assume that there is a fixed constant $K$ that uniformly bounds the switching curve from above. Our later approximations of the optimal policy are consistent with this assumption. Note that, as a consequence, $X_2(t) < K$ for all $t$. In the next construction of the fluid limit, we follow [12, Ch.9.6] and define

$$\begin{aligned} X_1^{(n)}(t) &= \frac{1}{n} X_1(nt), \\ X_2^{(n)}(t) &= X_2(nt), \end{aligned}$$

with initial condition $X_1(0) = n$. (Thus, the initial condition for the first component is different for each process in the sequence.) We will see shortly that the initial condition for $X_2$ is irrelevant. Note that for the first queue we scale both space and time, while for the second queue, which is uniformly bounded by the fixed constant $K$, we only scale time. Assuming that it exists, the fluid limit for the first queue is now defined as

$$x_1(t) = \lim_{n\to\infty} X_1^{(n)}(t).$$

Note that, for any fixed $t$, the random sequence $X_2^{(n)}(t)$ will converge weakly as $n \to \infty$, with the limiting distribution depending on (the value of the switching curve at) $x_1(t)$. Indeed, in the limit $n \to \infty$, $X_2^{(n)}(t)$ instantly reaches the stationary distribution [12, Ch.9.6] for each fixed $t$. In turn, the direction of $x_1(t)$ will depend on the distribution

$$\lim_{n\to\infty} \mathbb{P}\left(X_2^{(n)}(t) \leq x | X_1^{(n)}(t)\right),$$

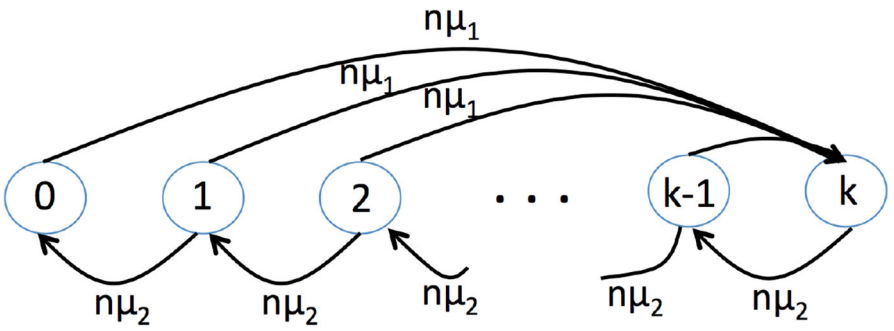


**Fig. 5** Graphical representation of $X_2^{(n)}(t)$

and in particular on its expectation, which we denote by

$$x_2(t) = \lim_{n\to\infty} \mathbb{E}\left[X_2^{(n)}(t)|X_1^{(n)}(t)\right].$$

Note that for all $n$, $\mathbb{E}\left[X_2^{(n)}(t)|X_1^{(n)}(t)\right]$ is random due to the randomness of $X_1^{(n)}(t)$, but in the limit it will be deterministic, since $x_1(t)$ is deterministic.

Recall that in the stochastic model, the optimal strategy is dictated by a switching curve $K(x)$ that gives the threshold value on the second queue for given $X_1(t)$. If a batch moves from the first to the second queue at time $t$, the process moves to the state $(y, K(y))$ on the switching curve, with $y$ such that $y + K(y) = X_1(t-) + X_2(t-)$. The size of the batch is $X_1(t-) - y$.

Let us now specify the local dynamics of the fluid limit. Since $K(\cdot)$ is assumed to be bounded, the size of the jump does not scale with $n$. Therefore, the fluid limit for the first component will not show these jumps. (See also the discussion in [3], where two different scalings are distinguished, one of which has a fluid limit with jumps and the other does not.) In the limit $n \to \infty$ the second component reaches stationarity instantly for any value of the first component, In addition, note that in the $n$-th system $(X_1^{(n)}(t), X_2^{(n)}(t))$, the rate with which batches move from the first to the second queue is $n\mu_1$. In the limit the process $x_1(t)$ will decrease continuously, the speed of movement being determined by the conditional distribution of the second component.

Let us now focus on $X_2^{(n)}(t)$, the size of the second queue in the $n$-th system, for a given level of $X_1^{(n)}(t)$ with constant threshold value $k$. For large $n$, $X_2^{(n)}(t)$ becomes a rapidly moving random variable with the stationary distribution of a batch-arrival queue in which the batches always lift the queue to the level $k$, as depicted in Fig. 5.

The stationary distribution (given the threshold value $k$) is therefore

$$\begin{aligned} \pi_i^{(k)} &= \pi_0^{(k)} \frac{\mu_1}{\mu_2} \left(\frac{\mu_1+\mu_2}{\mu_2}\right)^{i-1}, \quad i = 1, 2, \ldots, k, \\ \text{and } \pi_0^{(k)} &= \frac{1}{1 + \sum_{i=1}^{k} \frac{\mu_1}{\mu_2} \left(\frac{\mu_1+\mu_2}{\mu_2}\right)^{i-1}}. \end{aligned} \tag{4}$$

We will use the shorthand notation $\mathbb{E}[X_2|k]$ for the expectation of this distribution. The mean batch size is therefore $b(k) = k - \mathbb{E}[X_2|k]$.

This determines the dynamics of the first component in the fluid limit for a given limiting switching curve $k(x_1)$:

$$x_1'(t) = \lambda - b(k(x(t)))\mu_1,$$

as long as $x_1(t) > 0$, for a given arbitrary initial value $x_1(0)$. In our discussion above we took $X_1^{(n)}(0) = n$, to ensure that it is integer, which corresponds to $x_1(0) = 1$. The arguments remain valid for other positive values of $x_1(0)$ (for example by rounding the initial value of $nx_1(0)$ to an integer). In the next section we will exploit this description to determine a switching curve $k(\cdot)$ that approximates the optimal switching curve.

## 5 Fluid-based approximations of the optimal policy

We approximate the optimal strategy using two different approaches, both based on the fluid description in the previous section. The fluid model is used to approximate the trajectory of the stochastic process $X_1(t)$ by a smooth path. We emphasize that we *do not* formally work with the fluid limit, but instead directly use it to replace the stochastic process. The first method employs a fixed threshold strategy and the second approximation determines a dynamic threshold based on a greedy heuristic.

*Method 1*

In our first approach, we ignore the fact that we can adjust the threshold level over time. For any initial value $X_1(0) = x$ we approximate the threshold level $k = k(x)$ that minimizes the (approximated) cost until the first component is empty. We will denote the time at which this happens by $T = T(x)$. Replacing the stochastic path of $X_1(t)$ by the trajectory of the fluid model, and replacing $X_2(t)$ by its conditional expectation, we obtain

$$\min_k \left\{ c_1 \left( xT(x) + \frac{1}{2}(\lambda - b(k)\mu_1)T(x)^2 \right) + c_2 T(x)\mathbb{E}[X_2|k] \right\}. \tag{5}$$

To compute the threshold value $k$ that minimizes overall costs, we determine the time to empty the system:

$$x + (\lambda - b(k)\mu_1)T(x) = 0 T(x) = \frac{x}{b(k)\mu_1 - \lambda} \tag{6}$$

Equation (5) can thus be rewritten as

$$\min_k \left\{ c_1 x \frac{1}{2} T(x) + c_2 \mathbb{E}[X_2|k]\, T(x) \right\}. \tag{7}$$

From the stationary distribution in (4), we can numerically determine the value of $k$ that minimizes the approximated costs. Our first approximation thus replaces the optimal switching curve by a fixed-threshold strategy based on this value of $k$.

*Method 2*

The second method is based on a comparison of costs due to idleness in the second queue (implying loss of capacity if jobs from the first queue could have been moved earlier) and *additional* storage costs at the second queue (when jobs could have been transferred later from the first queue). These storage costs are therefore proportional to the number of jobs at the second queue.

*Loss of capacity*

Capacity loss is computed in the following manner. We again assume that the number of jobs in the first queue is large and that, at all times, queue 2 is in the equilibrium corresponding to the current threshold (say $k$, which is determined by queue 1). The maximum customer drain rate from the system per unit of time equals $\mu_2 - \lambda$. However, the effective outflow rate is lower than $\mu_2$, since it is interrupted when the second queue is empty. The fraction of time that the second queue is empty, that is $\pi_0^{(k)} = \mathbb{P}(X_2 = 0|k)$ in (4), is determined by the value of the threshold $(k)$. The actual outflow from the system is $\mu_2(1 - \pi_0^{(k)})$. Dividing the actual drain rate by the maximum drain rate gives the effective capacity per unit of time. The lost capacity can then be obtained as

$$1 - \frac{\mu_2\left(1 - \pi_0^{(k)}\right) - \lambda}{\mu_2 - \lambda} = \frac{\pi_0^{(k)}}{1 - \lambda/\mu_2}. \tag{8}$$

Since all jobs in the first queue will be delayed by this inefficiency, we obtain the total costs for capacity loss by multiplying (8) by the holding cost $c_1 X_1(t)$.

*Storage at queue 2*

The second component is intuitively easy. The average number of jobs waiting in the second queue is determined by the buffer level $k$. Each job faces an additional cost of $c_2 - c_1$ per time unit while being at queue 2, so total storage costs at the second queue are computed as $(c_2 - c_1)\mathbb{E}[X_2|k]$.

We combine the above into the following optimization problem:

$$\min_{k=k(x)} \left\{ c_1 x_1 \left( \frac{\pi_0^{(k)}}{1 - \lambda/\mu_2} \right) + (c_2 - c_1)\mathbb{E}[X_2|k] \right\}. \tag{9}$$

## 6 Experimental results of the fluid approximation

In this section, we demonstrate the accuracy of the fluid approximation methods proposed in Sect. 5. We compute the optimal switching strategies for several parameter choices by using the MDP solution and compare them to the proposed fluid approximation heuristics in terms of average costs and computation time.

In Fig. 6, we compare the asymptotics of the MDP solution and the two fluid approximations for two distinct parameter sets and for very large system states ($X_1(0) = 10^4$). From both figures, we observe that the greedy heuristic approximates the MDP threshold level very accurately, especially for a large number of jobs in the system. The fixed strategy consistently underestimates the switching curve, but does capture its shape quite well.

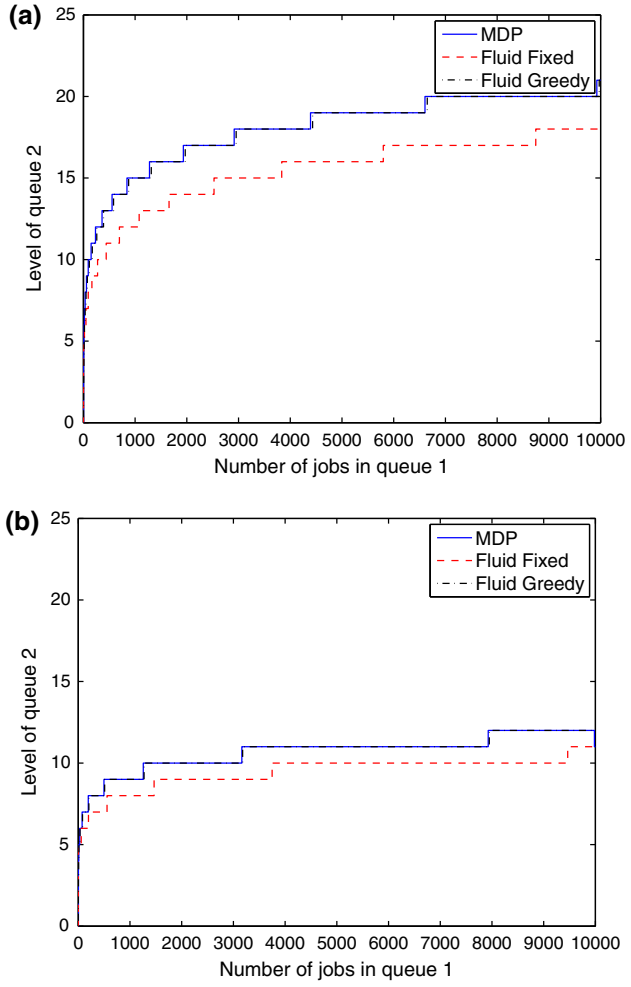


**Fig. 6** Comparison of the MDP result and the fluid heuristics for very large $n$. **a** Parameters $\lambda = 0.5$, $\mu_1 = 0.5$, $\mu_2 = 1.0$. **b** Parameters $\lambda = 0.9$, $\mu_1 = 1.5$, $\mu_2 = 1.0$

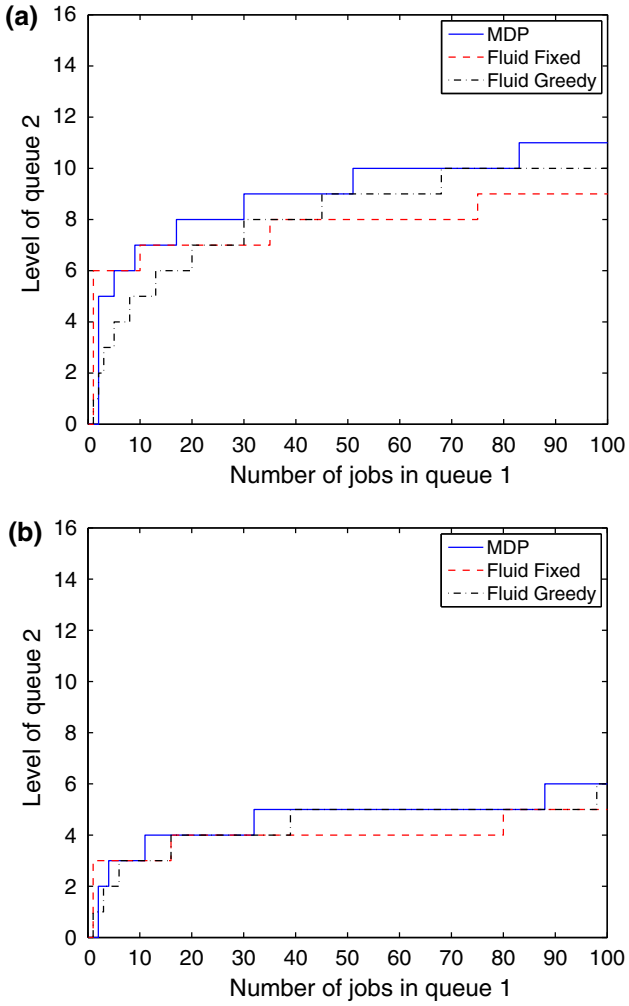


**Fig. 7** Comparison of the MDP result and the fluid heuristics for $n = 100$. **a** Parameters $\lambda = 0.7$, $\mu_1 = 0.5$, $\mu_2 = 1.0$. **b** Parameters $\lambda = 0.7$, $\mu_1 = 1.5$, $\mu_2 = 1.0$

In Fig. 7 we zoom in to lower levels, $n = 100$, giving a more detailed picture. We observe that close to the origin the "fixed" strategy overestimates the MDP curve for both parameter sets, while the "greedy" approach gives an underestimation. For smaller service rate at the first queue, the "greedy" heuristic is a worse approximation.

To gain more understanding of the accuracy of the approximations, we compare the average costs of the two fluid approximations with the optimal MDP solution. As a reference, we also compute the average cost for a fixed value threshold policy by using the Matrix-Geometric method which we have analysed in [15]. The chosen parameter values are those reported above in Table 1. Figure 8 shows the relative difference in average costs of the two fluid approximations and the fixed-threshold value from [15] with respect to the MDP solution. From these experiments, we see

**Table 1** Parameter sets used for the numerical experiments

| Set | Type | $\lambda$ | $\mu_1$ | $\mu_2$ | $c_1$ | $c_2$ |
|---|---|---|---|---|---|---|
| 1 | Batch | [0.1, 0.2, . . . , 0.9, 0.95] | 0.5 | 1.0 | 1 | 3 |
| 2 | Batch | [0.1, 0.2, . . . , 0.9, 0.95] | 1.0 | 1.0 | 1 | 3 |
| 3 | Batch | [0.1, 0.2, . . . , 0.9, 0.95] | 1.5 | 1.0 | 1 | 3 |

that the increase in average cost is relatively small. The "fluid greedy" heuristic shows the largest relative deviation of 8% on parameter set 1. In all others, the differences relative to the optimum are not more than a few per cent.

Figure 8 shows that on parameter set 1 the "greedy" approximation is much less accurate than the other two approximations. We already observed in the detailed graphs of Fig. 7 that a low service rate at the first queue causes a larger gap with the MDP threshold curve, particularly for small system states. This is reflected in the cost performance. For a large system load, the typical number of jobs in the system is larger, which reduces the impact of this underestimation of the "greedy" approach.

For all three parameter sets in Fig. 8, we observe that the greedy fluid approximation gives better results for heavy loads ($\rho \to 1$) than the other two approximations. The fixed fluid approximation appears to be the best all-round approximation.

To show the efficiency of the methods in terms of computation time, we averaged the computation times of the three parameter sets in Table 1 for increasing load. We separately investigate the time needed to compute the policy and the time it takes to compute the average costs of a given policy. The results are illustrated in Fig. 9.

The more relevant issue is the time needed to determine a good policy. Especially for heavily loaded systems, finding the optimal policy is computationally extremely demanding for the MDP method. Figure 9 shows that the computation times of the two fluid approximations are only mildly sensitive to the parameter choice, while the other methods quickly become slower for higher load. Note that the computation time for both fluid models is comparable, which explains the absence of the "fluid fixed" line in the figure. Even for small load, the fluid approximation is significantly faster than the MDP solution. We observe that the computation time depends on the load of the system and increases for higher load.

Although of less relevance, we also compared the time needed to compute the average cost of a given strategy using iterative approximation. (Note that for the MDP and the matrix-geometric approximation, the average cost is jointly determined with the policy itself. For the fluid approximations, these two phases are carried out separately.) It should be no surprise that for this metric all methods are essentially equivalent. It is quite likely that this computation time can be improved for all policies by using a more sophisticated computation scheme than direct iteration. Our goal here was to show that the differences are small.

## 7 The fluid model approximations with general service times

We continue our investigation of the fluid approximations and study their applicability under less restrictive assumptions on the service time distribution in the second queue,

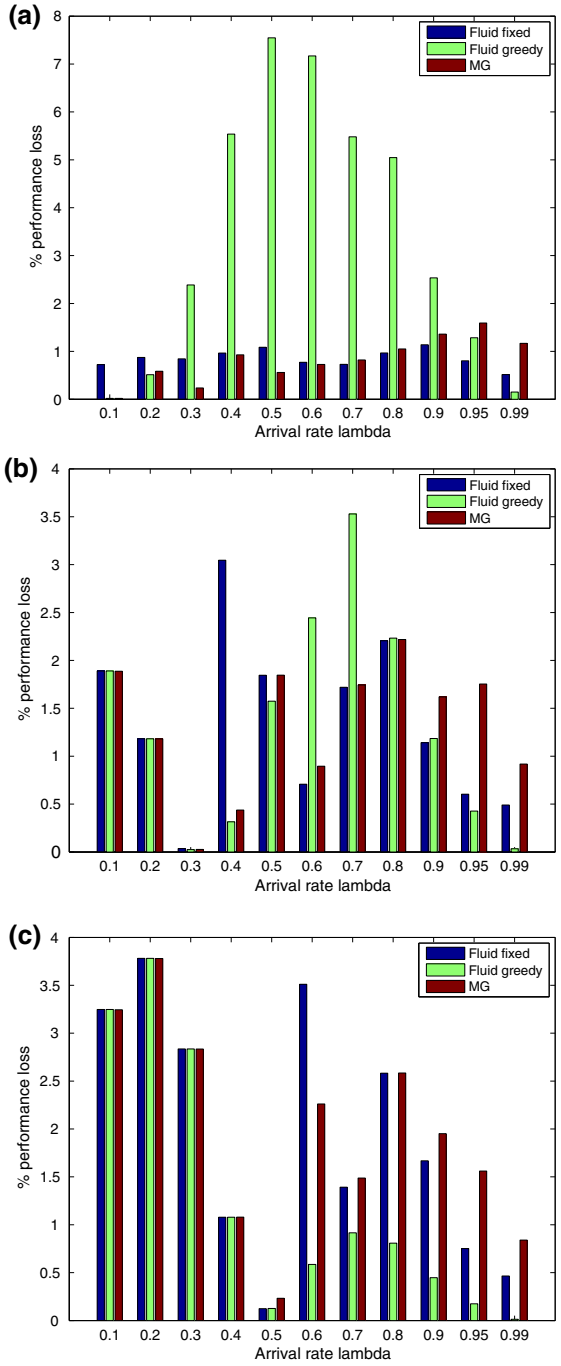


**Fig. 8** Comparison of the MDP result with both fluid approximations and the fixed-threshold approach of [15] for various parameter choices. **a** Parameter set 1. **b** Parameter set 2. **c** Parameter set 3

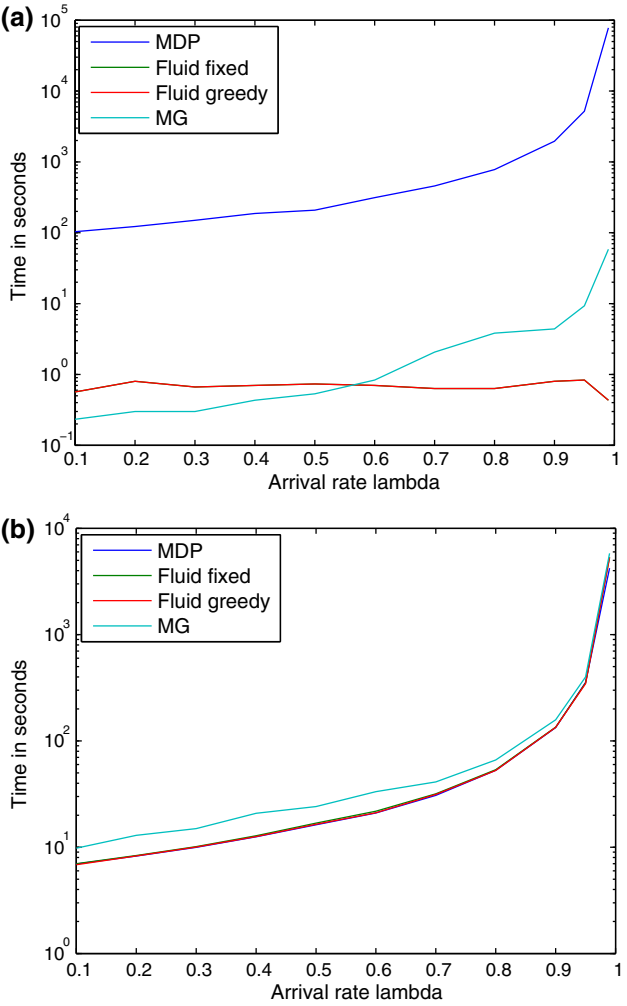


**Fig. 9** Time in seconds to compute the policy for increasing load of the system averaged over the service rate $\mu_1 \in [0.5, 1.0, 1.5]$ at the first queue. **a** Time to compute policy. **b** Time to compute average costs

which we now take to be of phase type. We concentrate on the single-service controlled tandem queue, which was also studied in [1,13,15]. The fluid approximations for the batch service model can also be used with phase-type services in the second queue, but solving the MDP for comparison becomes too demanding.

To apply the fluid approximations of Sect. 5 for the controllable tandem queue with two single server queues, we only need minor modifications: The second queue is now approximated with the usual $M/M/1/k$ queue instead of a batch-arrival queue, and we use its truncated geometric distribution as the conditional distribution for $X_2|k$. Since transfers are now all for single jobs, in the fluid formulation for the first queue we have a more limited control rule $b(\cdot) \in \{0, 1\}$.

**Table 2** Parameter sets for the single-service model with phase-type services in queue 2; $\lambda$ takes values in $\{0.7, 0.8, 0.9, 0.95\}$ and throughout we use $\mu_1 = 1.5$, $\mu_2 = 1.0$, $c_1 = 1$, $c_2 = 3$

| Set | Type | $v^2$ |
|---|---|---|
| 4 | Exponential | 1 |
| 5 | Erlang-2 | 1/2 |
| 6 | Erlang-4 | 1/4 |
| 7 | Erlang-6 | 1/6 |
| 8 | Hyper-2 | 2 |
| 9 | Hyper-2 | 4 |
| 10 | Hyper-2 | 6 |

Specifically, we will use the Erlang (with low variability) and the hyper-exponential (high variability) distributions for service durations in the second queue, and take the stationary distribution of the corresponding $M/PH/1/k$ queue as the conditional distribution for $X_2|k$. We keep the processing rates at the first and second queue ($\mu_1$ and $\mu_2$) fixed for all experiments, while adjusting the squared coefficient of variation. For an Erlang service distribution with $m$ phases, the coefficient of variation is given by

$$v^2 = \frac{1}{m}.$$

We parameterize the hyper-exponential distribution with two phases as follows:

$$F(x) = 1 - p_1 e^{-\nu_1 x} - p_2 e^{-\nu_2 x},$$

with $0 \leq p_1 = 1 - p_2 \leq 1$ and $\nu_1 > 0$, $\nu_2 > 0$. We use the method of "balanced means" to determine these parameters for a given mean $1/\mu_2$ and squared coefficient of variation $v^2$:

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{v^2 - 1}{v^2 + 1}}\right), \qquad p_2 = 1 - p_1, \qquad \nu_1 = 2p_1\mu_2, \qquad \nu_2 = 2p_2\mu_2.$$

As we will see, using the heuristic rules from the model with exponential services is straightforward, but the MDP solution suffers enormously in terms of computability, which demonstrates the need for approximations.

We extend our earlier experiments with the parameter sets presented in Table 2. We specify the service distribution at the second queue in the column "Type". To allow comparison between the different systems, we keep the average service duration at the second queue ($1/\mu_2$) fixed for all experiments and vary the coefficient of variation. In all our numerical experiments, the computations were performed by adequately truncating the state space, depending on the specific parameter values.

The results of this set of experiments are illustrated in Fig. 10 for Erlang service times of server 2, and for hyper-exponential services. We also show the corresponding

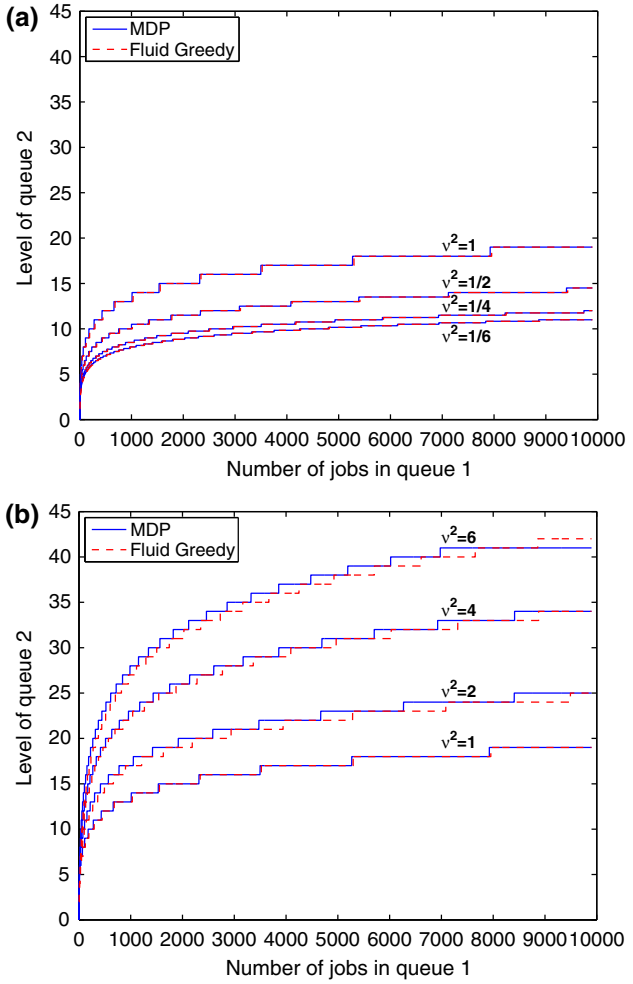


**Fig. 10** Comparison of the MDP results and fluid approximations for various $v^2$ in service variability at the second queue with load $\rho = 0.7$. **a** Erlang service in queue 2. **b** Hyper-exponential service in queue 2

graphs for exponential service durations (parameter set 4) as a reference. Clearly, the optimal switching curve obtained with MDP and the switching curve of the fluid approximation are again very close to each other. As might be expected, the switching curve is lower for less variable distributions (Erlang with many phases), because the departures from queue 2 can be predicted more accurately and thus there is less need to maintain a large buffer in queue 2. Similarly, for the hyper-exponential service durations with increasing variance, a more conservative strategy (larger threshold) is needed.

As before, we also investigate the accuracy in terms of achieving close to minimum cost. In Figs. 11 and 12 we observe that, as before, we obtain a better approximation in terms of cost for more highly loaded systems. This is natural, since the fluid approxi-

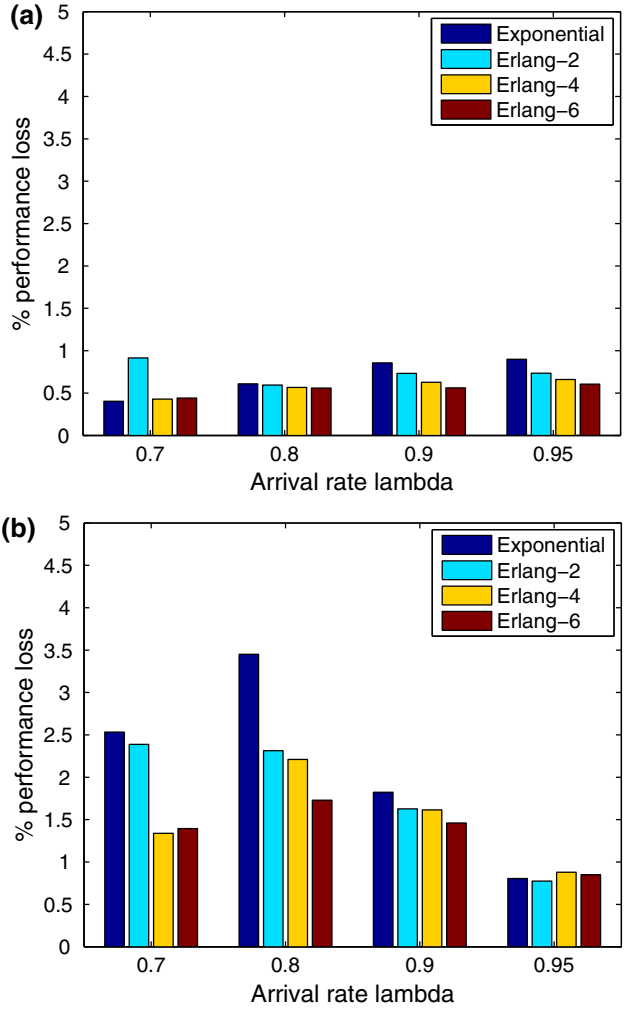


**Fig. 11** Performance results of the fluid approximations for parameter sets 4–7. **a** Fluid fixed. **b** Fluid greedy

mation is tailor made for states far from the origin. For hyper-exponential services at the second queue, we observe a better performance with the "greedy fluid" approach for larger coefficients of variation, while the "fixed fluid" approach does the opposite. This can be explained by the fact that larger states are more easily reached with more variability in the service times, which was better approximated by the greedy approach. For the lower variation of the Erlang service distributions, we see that both the "fluid fixed" and "greedy" approach give better performance when the coefficient of variation decreases. This suggests, unsurprisingly, that the fluid approximations are well suited for systems that have little variation in the service times.

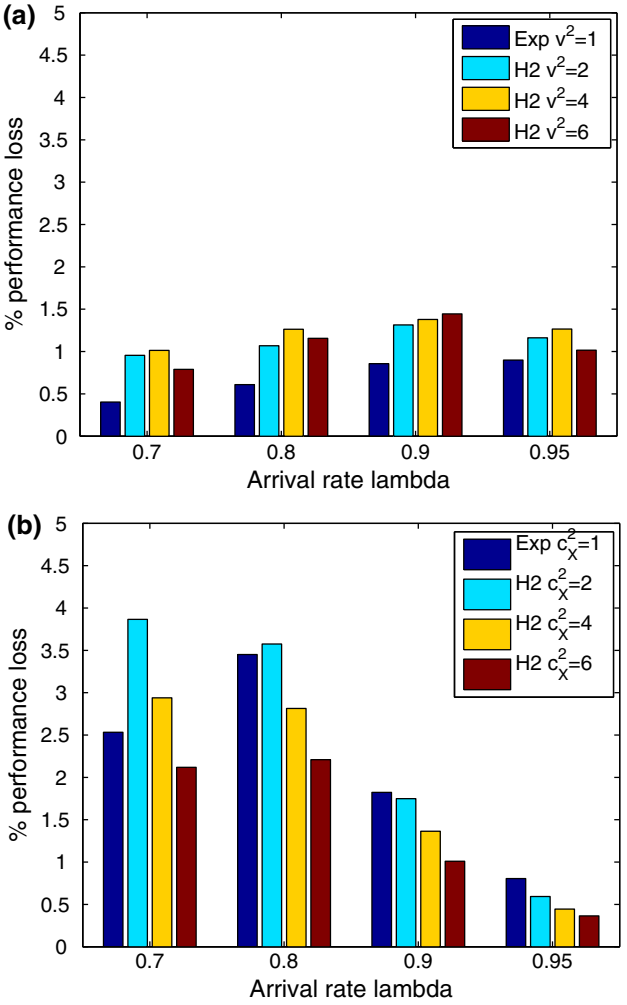


**Fig. 12** Performance results of the fluid approximations for parameter sets 8–10. **a** Fluid fixed. **b** Fluid greedy

## 8 Summary and outlook

We have investigated the structure of the optimal strategy to control the first service stage of a tandem queueing system with batch services. In the Markovian setting, we formulated an MDP to determine the optimal strategy in terms of *when to serve at the first stage* and *how large the batch size should be*. Solving the MDP numerically is extremely computationally intensive. To gain more understanding of the shape of the optimal MDP policy, we developed approximations and computationally efficient heuristics that are very close to the MDP strategy, especially for high loads.

For the design of our heuristics we noted that the optimal MDP strategy is characterized by a switching curve that is rather flat. In order to formulate a meaningful approximating fluid model, we applied different scalings to the two queues, cf. the approach in [12, Ch.9.6]. To the best of our knowledge, this has not been applied to stochastic control problems before.

We have developed two different heuristics based on the fluid model approximation. The "fixed fluid" heuristic underestimates the optimal MDP strategy in states with a large number of jobs in the system, while the "greedy fluid" approach follows the optimal MDP switching curve quite closely. The average costs of the "fixed fluid" approach remain within a few per cent of the average costs of the MDP solution for a wide range of parameters. The "greedy" approach becomes more accurate for higher load.

Encouraged by the simplicity and the accuracy of the two approximations for the batch tandem system, we investigated the applicability for non-exponential service durations in the second queue. For the batch service model, the MDP formulation quickly becomes numerically intractable, leaving us with no bench mark to test our approximations. For this reason, we illustrated the potential of the approach for more general service times by only allowing single services at the first station. We again obtain an approximation function that closely follows the optimal MDP policy. As before, the accuracy of the "greedy fluid" approximation improves for increasing load and the "fixed fluid" approach performs well for a wide range of parameters.

The proposed fluid approach is computationally very fast. Solutions are available within a second (evaluated on a Macbook Pro, dated from 2013 with 8 GB internal memory). This suggests that the approach is worth exploring for larger queueing networks with non-exponential service times.

## References

1. Avram, F.: Optimal control of fluid limits of queueing networks and stochasticity corrections. Lect. Appl. Math. **33**, 1–36 (1997)
2. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton (2003)
3. Bortolussi, L., Tribastone, M.: Fluid limits of queueing networks with batches. In: Proceedings of 3rd ACM/SPEC International Conference on Performance Engineering (2012)
4. Foss, S., Kovalevskii, A.: A stability criterion via fluid limits and its application to a polling model. Queueing Syst. **32**, 131168 (1999)
5. Koole, G.: Convexity in tandem queues. Probab. Eng. Inf. Sci. **18**(01), 13–31 (2004)
6. Kurtz, T.G.: Solutions of ordinary differential equations as limits of pure Markov processes. J. Appl. Prob. **7**, 4958 (1970)
7. Larrañaga, M., Boxma, O.J., Núñez Queija, R., Squillante, M.S.: Efficient content delivery in the presence of impatient jobs. In: Teletraffic Congress (ITC 27), 2015 27th International. IEEE, (2015)
8. Lippman, S.A.: Applying a new device in the optimisation of exponential queueing systems. Oper. Res. **23**, 687–710 (1975)
9. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, Hoboken (1994)
10. Rajat, D.K., Serfozo, R.F.: Optimal control of batch service queues. Adv. Appl. Probab. **5**(2), 340–361 (1973)
11. Rajat, D.K.: Optimal control of batch service queues with switching costs. Adv. Appl. Probab. **8**(1), 177–194 (1976)
12. Robert, P.: Stochastic Networks and Queues, vol. 52. Springer, Berlin (2013)

13. Rosberg, Z., Varaiya, P.P., Walrand, J.: Optimal control of service in tandem queues. IEEE Trans. Autom. Control **27**(3), 600 (1982)
14. Silva, D.F., Zhang, B., Ayhan, H.: Optimal admission control for tandem loss systems with two stations. Oper. Res. Lett. **41**(4), 351–356 (2013)
15. van Leeuwen, D., Núñez-Queija, R.: Near-Optimal Switching Strategies for a Tandem Queue. In: Boucherie, R.J., van Dijk, N.M. (eds.) Markov Decision Processes in Practice, pp. 439–459. Springer, Berlin (2017)
16. Veatch, M.H., Lawrence, M.W.: Optimal control of a two-station tandem production/inventory system. Oper. Res. **42**(2), 337–350 (1994)
17. Weber, R.R., Stidham, S.: Optimal control of service rates in networks of queues. Adv. Appl. Probab. **19**, 202–218 (1987)
18. Zhang, R., Phillis, Y.A.: Fuzzy control of arrivals to tandem queues with two stations. IEEE Trans. Fuzzy Syst. **7**(3), 361–367 (1999)