Optimal steady-state and transient trajectories of a two queue switching server

Dirk van Zwieten[*]
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
D.A.J.v.Zwieten@tue.nl

Erjen Lefeber
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
A.A.J.Lefeber@tue.nl

Ivo Adan
Eindhoven University of Technology
P.O. Box 513, 5600 MB
Eindhoven, the Netherlands
IAdan@tue.nl

ABSTRACT

We consider two fluid queues attended by a switching server and address the optimal steady-state and transient trajectory problems. The steady-state problem is formulated as a quadratic problem (QP), given a fixed cycle time. Evaluating the QP problem over a range of cycle times results in the optimal steady-state trajectory. We minimize the holding costs, backlog costs and setup costs, allow setup times and allow constraints on queue contents, cycle times and service times. Second, given initial conditions, we derive the optimal transient trajectory that leads to the optimal steady-state trajectory in a finite amount of time with minimal costs. The transient switching behavior and optimal initial modes are also addressed.

Categories and Subject Descriptors

G.1.6 [**Optimization**]: Quadratic programming methods

General Terms

Performance, Quadratic programming, Steady-state trajectory, Transient trajectory.

1. INTRODUCTION

The control of systems with switching behavior is a problem of great importance and even the most simple system, a server attending two queues, has been investigated by many researchers, see for example [1, 2, 3, 4, 5, 6, 8, 9, 11]. We consider a system of two fluid queues. Both queues share a single server which can serve only one queue at a time. Switching service to another queue might take time and/or involves switching costs. Such models arise in numerous contexts, such as manufacturing systems, signalized traffic intersections, computer communication networks and hospital rooms. Optimal schedules for manufacturing systems can reduce costs via, for instance, lowering the required storage capacity and shortening lead times. For traffic signals at signalized intersections, optimal schedules can reduce congestion, and thereby reduce the amount of environmentally harmful emissions, and improve mobility.

The optimal scheduling problem can be divided into two subproblems. The first problem is the derivation of optimal steady-state trajectories. The second problem concerns the determination of optimal transient trajectories, that is, trajectories that lead to the optimal steady-state trajectory in finite time and at minimal costs. Both subproblems have been intensively studied, see [1, 2, 3, 4, 5, 6, 8, 9, 11] and references therein. In most work, systems are restricted. Setup times, setup costs, backlog or limited queue contents are required, omitted or not allowed. Also, often the system is studied under the simplifying condition that the system is symmetric, see [1, 2, 6, 9]. In this work, we present a method to derive the optimal steady-state trajectory for a two queue switching server without restrictions on parameters and with the flexibility of allowing setup times, setup costs and backlog, as well as constraints on cycle time, service time and queue contents.

We follow the general framework introduced by [7] and model the production flow as continuous rather than discrete. Similar to [12], we formulate the steady-state problem as a QP with the addition of backlog and setup costs.

Once the optimal steady-state trajectory is known, we study the best way of reaching it from any initial state, i.e., with minimal costs. This is a transient optimization problem, occurring for instance in case of a machine which is failure prone, or in case of a traffic intersection which gives priority to busses [13]. In these cases, we assume that deviations from the steady-state trajectory rarely occur, allowing the system to recover to the steady-state situation after each interruption. We present optimal transient trajectories for systems without backlog and, for systems without capacity constraints, the policy for optimal transient behavior.

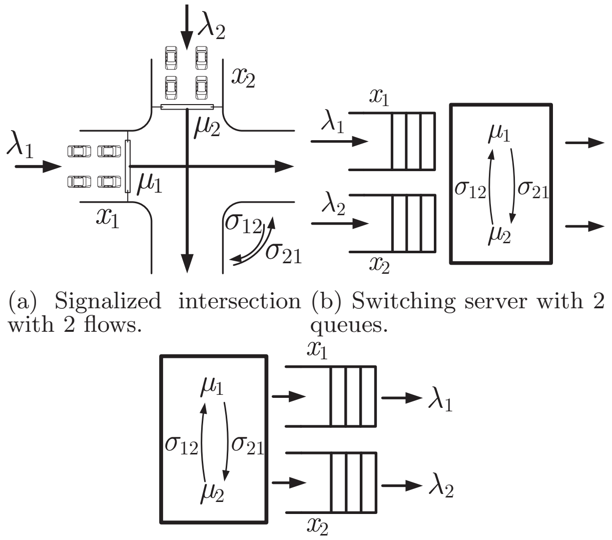
The remainder of this paper is organized as follows. Section 2 describes the system and presents the constraints. The optimal steady-state problem is addressed in Section 3 and examples of optimal trajectories are presented. In Section 4, the optimal transient problem is addressed. Conclusions are provided in Section 5.

[*] Corresponding author.

VALUETOOLS 2013, December 10-12, Torino, Italy
Copyright © 2014 ICST 978-1-936968-48-0
DOI 10.4108/icst.valuetools.2013.254375

2. SYSTEM DESCRIPTION

We consider a system of two queues served by a single switching server. Fluid arrives at each queue $i = 1, 2$ with arrival rate λ_i. The content of queue i at time t is denoted by $x_i(t)$. The server is limited to only serve one queue at a time. If the server serves queue i, the service rate is given by $r_i \in [0, \mu_i]$. In [10] it is shown that for optimal policies, a server, once serving a queue, does not idle and serves at *maximal* rate. Three examples of the system under consideration are presented in Figure 1, a signalized traffic intersection with two flows in Figure 1a, a 2-queue switching server in Figure 1b and a 2-product manufacturing system in Figure 1c. The latter system has constant demands λ_i instead of constant arrivals.



(a) Signalized intersection with 2 flows.

(b) Switching server with 2 queues.

(c) Manufacturing system with 2 jobs.

Figure 1: Different two queue switching server layouts.

Define the load of queue i by $\rho_i = \frac{\lambda_i}{\mu_i}$. For stability, i.e., all arriving fluid can be served, it is required that $\rho_1 + \rho_2 < 1$.

Typically, switching service between different queues implies a setup process, either a *setup time* σ_{ij} for switching from queue i to queue j, *setup costs* s_{ij} or a combination of these. For instance, a setup time can be reserved for vehicles to leave the intersection after the queue has received a red light (end of service), thereby preventing collisions, or for a machine to adjust configurations to do some cleaning.

Given the setup times and cyclic behavior, we assume that the system can operate in four modes, denoted by $m \in \{1, 2, 3, 4\}$. Without loss of generality, the first mode, $m = 1$, indicates a setup to serve queue 1, $m = 2$ indicates serving queue 1, $m = 3$ indicates a setup to serve queue 2 and $m = 4$ indicates serving queue 2. Note that for a system without setup times, i.e., $\sigma_{ij} = \sigma_{ji} = 0$, modes 1 and 3 have a duration of zero time units.

A service (idle) time is defined as the uninterrupted interval during which the queue is (not) served. Note that during a setup no queue content is processed, hence a setup is part of the idle time. The duration of a service (idle) time for queue i is nonnegative and is denoted by τ_n (t_n). The service time is divided into two parts,

$$\tau_n = \tau_n^{\mu} + \tau_n^{\lambda}.$$

where the duration of serving at maximal rate is indicated by τ_i^{μ} and the duration of serving at arrival rate by τ_i^{λ}. Note that serving at arrival rate, i.e., $\tau_n^{\lambda} > 0$, occurs only if both backlog is not allowed and the queue is empty. This duration is referred to as *slow-mode*, since capacity is wasted. However, as indicated in [11] and shown in Section 3.1, using a slow-mode might lead to optimal behavior, since it enlarges the cycle time and thereby reduces the fraction of time spent on setups, which also wastes capacity.

The *cycle time* T is the time it takes to serve both queues in a cycle. The cycle time consists of setup times and times serving the queues, i.e.,

$$T = \sigma_{21} + \tau_1^{\mu} + \tau_1^{\lambda} + \sigma_{12} + \tau_2^{\mu} + \tau_2^{\lambda}. \tag{1}$$

2.1 System constraints

Depending on the system under consideration, some restrictions can be imposed. As presented below, these constraints can originate from operational or safety issues. Note that these constraints are not mandatory, but can be included if required. Cycle time constraints can originate from, for instance, limiting the cycle time of a manufacturing system to the operator's available time or requiring a minimal cycle time for safety reasons in traffic intersections. Therefore, minimal and maximal cycle times, respectively $T^{\min}$ and $T^{\max}$, can be taken into account,

$$T^{\min} \le T \le T^{\max}. \tag{2}$$

Furthermore, bounds on service times, denoted by $\tau_i^{\min}$ and $\tau_i^{\max}$, can be required, e.g., minimal and maximal service (green) times for traffic intersections. The service time constraints are imposed via

$$\tau_i^{\min} \le \tau_i \le \tau_i^{\max} \qquad i = 1, 2. \tag{3}$$

In addition to the constraints on cycle and service times, the queue lengths can be bounded, e.g., finite buffer capacity, given by

$$x_i(t) \le x_i^{\max} \qquad i = 1, 2. \tag{4}$$

Furthermore, for stability, all arrivals must be served within a period, resulting in

$$\lambda_i T = \mu_i \tau_i^{\mu} + \lambda_i \tau_i^{\lambda} \qquad i = 1, 2. \tag{5}$$

Note that we can impose additional constraints, regarding service times and/or queue contents. Given the system description and the constraints, we present a method to derive the optimal steady-state trajectory in Section 3. This trajectory is used in Section 4 to derive the optimal transient trajectory.

3. STEADY-STATE TRAJECTORY

Multiple performance criteria exists for evaluating the trajectory. For the system under consideration, cycle time, flow-time or costs are commonly used criteria. In this paper, we focus on minimizing the costs. However, other criteria can be easily incorporated. In the remainder of this paper we assume linear costs on the queue levels. Inventory costs c_i^+ are proportional with $x_i^+(t)$, where $x_i^+(t) = \max(x_i(t), 0)$. Backlog costs c_i^-, which for instance arise when production is behind on the demand for the system depicted in Figure 1c, are proportional with $x_i^-(t)$, where $x_i^-(t) = \min(x_i(t), 0)$. This results in the following costs for the optimal steady-state trajectory

$$J_s = \frac{1}{T}\int_0^T c_1^+ x_1^+(\tau) + c_1^- x_1^-(\tau) + c_2^+ x_2^+(\tau) + c_2^- x_2^-(\tau) d\tau ... \\ + \frac{s_{21} + s_{12}}{T}.$$

The trajectory minimizing J_s is the optimal steady-state trajectory. The optimal trajectory is a trade-off between loss of capacity due to setups, slow-modes and the average setup costs. Elongating the cycle time by including a slow-mode or creating backlog, results in less switches over time where capacity is lost due to setups and lowers the average setup costs.

For a system without both setup costs and backlog, i.e., $s_{21} = s_{12} = 0$ and $x_i(t) \geq 0$, the optimal steady-state trajectory can be analytically derived from [11]. This trajectory is depicted in Figure 2, and contains at most one slow-mode, i.e., $F = A$ and/or $C = D$. If no setup times are considered, $D = E$ and $A = B$. The optimal policy consists of serving queue i until the other queue j reaches a threshold. Therefore, the trajectory can include slow-modes at both queues. A special case of this model, with $\mu_1 = \mu_2$ and $c_1 = c_2$ has been studied in [1, 2, 5, 6, 9] and it is shown that the optimal policy is a *clearing* policy, i.e., the server empties a queue and then switches to serve the other queue.

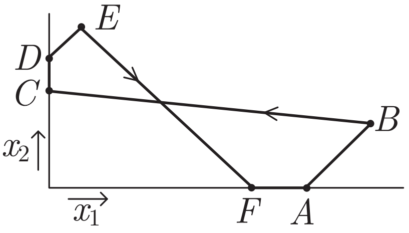


Figure 2: Optimal steady-state trajectory, with characteristic points $A - F$.

We denote the minimal queue content of queue i by $\underline{x}_i$. The optimal trajectory implies that the contents of both queues are zero at least once in a cycle. Otherwise, a constant inventory or backlog is present which results in non-optimal behavior. Therefore, $\underline{x}_i \leq 0$, $i = 1, 2$.

Similar to the approach in [12], where the optimal steady-state trajectory for a switching server that competes over multiple queues (without setup costs and backlog) is presented, we derive the optimal steady-state trajectory using quadratic programming. For a fixed cycle time, the optimization problem is quadratic and by evaluating the performance of the system over a range of cycle times the optimal trajectory is derived. Moreover, we allow setup costs and backlog in our approach. Total inventory and backlog of a queue during a cycle can be derived regarding the service times, due to the fluid flows and cyclic behavior. Figure 3 presents the contents of queue i during a single cycle. All service and setup times are indicated, together with the slope rates.

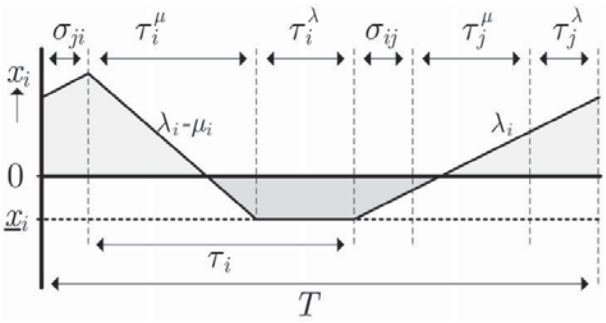


Figure 3: Evolution of x_i during a cycle, including setup and service periods and rates of increase/decrease.

Total inventory of queue i during a cycle is denoted by w_i^+ and total backlog by w_i^-:

$$w_i^+ = \int_0^T x_i^+(\tau)d\tau = \frac{1}{2}\frac{\lambda_i\mu_i}{\mu_i - \lambda_i}(T - \tau_i)^2 - \underline{x}_i T + w_i^-, i = 1, 2, \tag{6a}$$

$$w_i^- = \int_0^T x_i^-(\tau)d\tau = \frac{\underline{x}_i^2}{2}(\frac{1}{\mu_i - \lambda_i} + \frac{1}{\lambda_i}) + \tau_i^\lambda \underline{x}_i, \qquad i = 1, 2. \tag{6b}$$

It can be easily seen that the expressions for w_i^+ and w_i^- are quadratic in the optimization variables τ_i and $\underline{x}_i$. In case of two queues, where every queue is served only once in a cycle, the setup costs are constant and do not influence the optimal steady-state trajectory for a fixed cycle time. However, these costs are important for the time-average costs.

The optimal steady-state costs, for a fixed cycle time, are given by

$$J_s(T) = \frac{1}{T}\left(s_{12} + s_{21} + Q_s(T)\right), \tag{7}$$

with $Q_s(T)$ the solution to the quadratic programming problem given by

$$Q_s(T) = \min_{\tau_i^\mu, \tau_i^\lambda, \underline{x}_i} \sum_{i=1}^{2} c_i^+ w_i^+ + c_i^- w_i^-, \tag{8a}$$

$$s.t. \quad \tau_i^{\min} \leq \tau_i \leq \tau_i^{\max}, \qquad i = 1, 2, \tag{8b}$$

$$\underline{x}_i \leq x_i^{\max} - (\mu_i - \lambda_i)\tau_i^\mu, \qquad i = 1, 2, \tag{8c}$$

$$\lambda_i T = \mu_i \tau_i^\mu + \lambda_i \tau_i^\lambda, \qquad i = 1, 2, \tag{8d}$$

$$T = \sigma_{21} + \tau_1^\mu + \tau_1^\lambda + \sigma_{12} + \tau_2^\mu + \tau_2^\lambda, \tag{8e}$$

where (8c) follows from (4), i.e., the maximal queue content is given by $\underline{x}_i + (\mu_i - \lambda_i)\tau_i^\mu$, as can be seen in Figure 3. Then, the solution $J_s(T)$ with minimal costs for cycle times within the range (2) renders the optimal steady-state costs J_s^*. From the optimal service times and optimal minimal queue contents, belonging to J_s^*, the contents of each queue during a cycle can be derived. Combining the contents of both queues then results in the optimal steady-state trajectory.

Note that this approach can be easily extended to a system with > 2 queues, provided that a fixed service sequence is given. Otherwise, evaluating all possible service sequences, and selecting the sequence with the best results, results in the optimal trajectory. Consider, for instance, a system with three queues. Then, if each queue is served once in a cycle, two different service sequences exist, i.e., serving queues 1, 2 and 3 consecutively, or serving queues 1, 3 and 2 consecutively. Note that if all setup times and costs are identical and each queue is served only once in a cycle, the order of queues in a cycle is irrelevant, as each order of queues has identical costs. Next, for each sequence, the optimal steady-state costs are derived, with similar costs as in (7), total inventory and backlog (6) and constraints (1), (3)-(5). Then, the sequence with lowest steady-state costs renders the optimal sequence. If a queue is served multiple times in a cycle, e.g., serving queues 1, 2, 1 and 3 consecutively, the total inventory and backlog as presented in (6) does not hold for the queue(s) that is (are) served multiple times. These levels can be derived, in a similar way as presented above, but are omitted since it is outside the scope of this paper. The interested reader for the optimal steady-state trajectory of a switching server that competes over multiple queues without setup costs and backlog, and that can also serve multiple queues simultaneously at a queue-dependent rate, is referred to [12].

3.1 Examples

Using the method described above, we illustrate some optimal steady-state trajectories for this system. We start from a simple system, without setup costs, constraints on cycle and service times and no backlog, and then add more parameters and restrictions. For a non-symmetric system, with parameters

$$\lambda_1 = 2, \qquad \lambda_2 = 1, \tag{9a}$$

$$\mu_1 = 8, \qquad \mu_2 = 4, \tag{9b}$$

$$\sigma_{21} = 3, \qquad \sigma_{12} = 7, \tag{9c}$$

$$c_1^+ = 8, \qquad c_2^+ = 1, \tag{9d}$$

the optimal steady state costs $J_s^* = 120$ is reached for a cycle time of $T = 32$. Corresponding service times are $\tau_1^\mu = 6$, $\tau_1^\lambda = 8$, $\tau_2^\mu = 6$ and $\tau_2^\lambda = 0$. The corresponding queue contents during a cycle are depicted in Figure 4. Note that the minimal cycle time for this system to be stable is 20 time units.

The optimal steady-state trajectory is depicted in Figure 5a. The trajectory, which includes a slow-mode, can also be derived analytically using [11]. Adding setup costs $s_{21} = 300$ and $s_{12} = 200$ to the system results in the optimal trajectory depicted in Figure 5b. It can be seen that this addition elongates the cycle time and increases the duration of the slow-mode. Allowing backlog, with $c_1^- = 50$ and $c_2^- = 3$, shifts the optimal trajectory downwards and enlarges the cycle time, see Figure 5c. Note that no backlog occurs at queue 1, which is optimal due to the long slow-mode. Next, the service time of queue 1 is restricted ($\tau_1^{\max} = 15$), resulting in the optimal trajectory depicted in Figure 5d. Adding upper bounds on the queue contents $x_1^{\max} = 35$ and $x_2^{\max} = 16$, also reduces the cycle time, see Figure 5e. In Figure 5f the optimal trajectory of the system with a maximal cycle time $T^{\max} = 25$ is depicted. This trajectory has a cycle time of 20 time units, the minimal required cycle time (no slow-modes), and also queue 1 has backlog. Finally, in Figure 5g the optimal steady-state trajectory for this system without setup times is depicted. Due to the setup costs, this trajectory is not the fixed point $(0, 0)$.
The optimal steady-state trajectory is used for deriving the optimal transient trajectory, presented below.

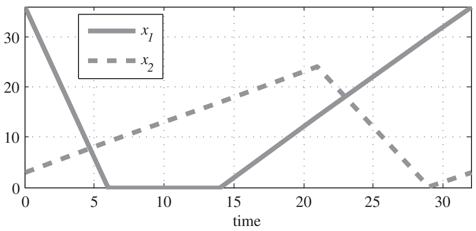


Figure 4: Optimal queue contents during a cycle for the system with parameters (9).

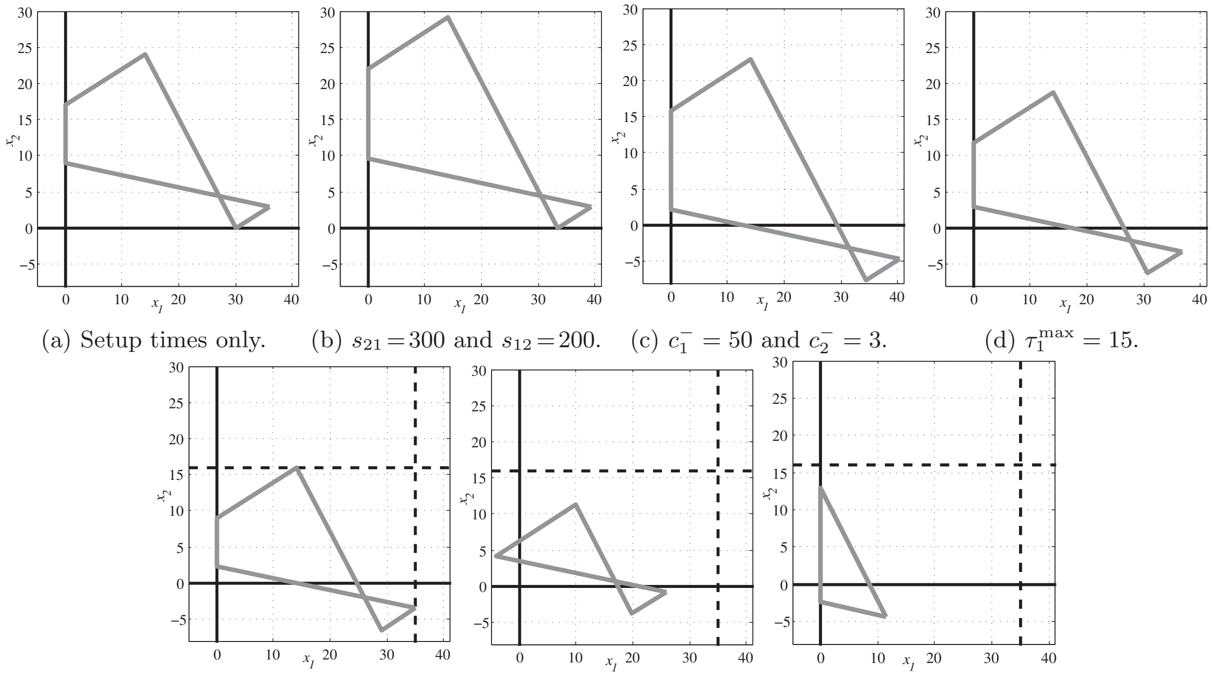
4. OPTIMAL TRANSIENT TRAJECTORY

The transient optimization problem is that of bringing the system back to the optimal steady-state trajectory at minimal costs. Machine failure in a manufacturing application or bus priorities in a signalized traffic intersection are two examples that remove the system from the steady-state trajectory. We assume that deviations from the steady-state trajectory rarely occur, allowing the system to recover to the steady-state situation after each interruption. A transient solution is defined as a trajectory in the $x_1 - x_2$ space that leads to the optimal steady-state trajectory in a *finite* amount of time. An *optimal* transient solution is a transient solution which minimizes the costs of reaching the optimal steady-state trajectory. In the remainder, backlog is not allowed for the transient trajectory, i.e., $x_i(t) \geq 0$, $i = 1, 2$. Therefore, if queue i is served, the service rate is given by

$$r_i(t) = \begin{cases} \mu_i & \text{if } x_i(t) > 0, \\ \lambda_i & \text{if } x_i(t) = 0. \end{cases}$$

In other words, if the queue is nonempty, service is at maximal rate, otherwise at arrival rate. Note that backlog might be incorporated in a similar way as presented for the steady-state trajectory, i.e., by deriving the total inventory and total backlog during a cycle. However, as both queues in the transient trajectory are not necessarily zero at least once during a cycle, this derivation is more complex. Therefore, solving the transient problem for systems with backlog might result in using other optimization methods as QP. For the system without backlog, the transient costs are defined by

$$J_p = \liminf_{t \to \infty} \int_0^t c_1^+ x_1(\tau) + c_2^+ x_2(\tau) + s_{21} n_1(\tau) + s_{12} n_2(\tau) - J_s^* d\tau, \tag{10}$$



(a) Setup times only. (b) $s_{21}=300$ and $s_{12}=200$. (c) $c_1^- = 50$ and $c_2^- = 3$. (d) $\tau_1^{\max} = 15$.

(e) $x^{\max} = [35, 16]^\top$. (f) $T^{\max} = 25$. (g) No setup times.

Figure 5: Optimal steady-state trajectories for the system with parameters (9). The system with setup times only (a) is extended with setup costs (b), backlog (c), maximal service time (d), maximal queue contents (e) and maximal cycle time (f). In (g), no setup times are considered.

with $n_i(t) = \frac{1}{\sigma_{ji}}$, $j \neq i$ during a setup to queue i and $n_i(t) = 0$ otherwise. For the remainder of this paper, we define the initial state by $[m(0), x(0)^\top] \in \{1,2,3,4\} \times \mathbb{R}_+^2$, $\mathbb{R}_+ := [0, \infty)$, as the state immediately after removal from the periodic solution, e.g., after the machine failure or bus priority. In order to reach the steady-state trajectory from every possible initial state in finite time, the steady-state trajectory requires a slow-mode, since serving at a lower rate, i.e., not at full capacity, provides the transient trajectory to 'catch up' with the steady-state trajectory.

For a *fixed* number of cycles C, we present the transient optimization problem as a QP problem. A cycle, starting at mode m, is defined as the series of operations until the end of the previous mode (which is 4 for mode 1, 1 for mode 2, etc.). Denote by $\tau_{i,c}$ the service time of queue i for the c-th cycle ($c \leq C$), consisting of the service time at maximal rate $\tau_{i,c}^{\mu}$ and the service time at arrival rate $\tau_{i,c}^{\lambda}$. In the remainder of this paper we assume that the initial mode is 1, i.e., start setting up to serve queue 1, and derive the QP problem for this particular case. The QP problems for the other initial modes can be derived similarly. Constraints for the transient problem (c.f. (2)-(4)) are listed below. Minimal and maximal cycle time constraints are:

$$\begin{aligned} T^{min} \leq& \sigma_{21} + \tau_{1,c}^{\mu} + \tau_{1,c}^{\lambda} + \sigma_{12} + \tau_{2,c}^{\mu} + \tau_{2,c}^{\lambda} \leq T^{max}, \\ & c = 1, ..., C. \end{aligned} \tag{11a}$$

Minimal and maximal service times:

$$\tau_i^{min} \leq \tau_{i,c} \leq \tau_i^{max}, \qquad \text{for } i \in 1,2,\ c = 1, ..., C. \tag{11b}$$

Maximal queue contents:

$$x_i(t) \leq x_i^{\max}, \qquad \text{for } i \in 1, 2. \tag{11c}$$

Considering the transient optimization problem for an infinite number of cycles, the transient trajectory would remain on the steady-state trajectory once it is reached. However, due to the finite number of cycles considered in the QP problem, a termination effect occurs, e.g., elongating the cycle time and/or lowering the accumulated queue contents in the final cycle (or even earlier) can lower the costs. Therefore, we consider in the QP problem $C > 2$ cycles and require the $C-2$th cycle of the transient trajectory to follow the optimal steady-state trajectory, i.e., the queue contents of the $C-2$th cycle of the transient trajectory are identical to the queue contents during the corresponding cycle of the optimal steady-state trajectory. If so, the trajectory is defined as a *feasible* transient trajectory, otherwise the trajectory is *infeasible*. Then, for a feasible trajectory, the final two cycles, which can include termination effects, are not considered.

As an illustrative example, the queue levels of a transient trajectory with $C = 5$ is presented in Figure 6. The setup and service periods are also indicated. Since the trajectory starts in mode 1, cycle c starts in mode 1 and ends at the end of mode 4, i.e., serving at rate $\tau_{2,c}^{\lambda}$. The transient trajectory converges to the steady-state trajectory, presented in gray, during service of queue 1 at arrival rate in the third cycle, i.e., during $\tau_{1,3}^{\lambda}$. The trajectory remains on the steady-state trajectory until the fifth, and final, cycle. In the final cycle, service times are zero and the steady-state trajectory is left, reflecting the termination effect. Note that the depicted trajectory is not the optimal transient trajectory, as for instance slow-modes $\tau^\lambda_{1,1}$ and $\tau^\lambda_{2,1}$ occur while the queues are non-empty. Also, the trajectory is not feasible, as the third cycle not entirely follows the optimal steady-state trajectory.

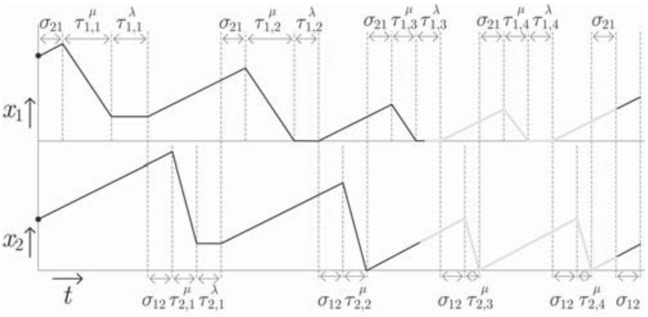


Figure 6: Queue levels during transient phase ($C = 5$), convergence to steady-state trajectory during $\tau^\lambda_{1,3}$.

Let us denote by $x_{i,c}$ the content of queue i at the end of the c-th cycle:

$$x_{i,c+1} = x_{i,c} + \lambda_i(\sigma_{12} + \sigma_{21} + \tau^\mu_{j,c} + \tau^\lambda_{j,c}) - (\mu_i - \lambda_i)\tau^\mu_{i,c},$$
$$i = 1, 2, \quad c = 1, ..., C, \tag{12}$$

where $x_{i,0} = x_i(0)$. Then, the total queue contents $w_{i,c}$ of queue i for cycle c can be derived by:

$$\begin{aligned} w_{1,c} =&(x_{1,c-1} + \tfrac{1}{2}\lambda_1\sigma_{21})\sigma_{21} + (x_{1,c-1} + \lambda_1\sigma_{21} - ... \\ &\tfrac{1}{2}\mu_1\tau^\mu_{1,c})\tau^\mu_{1,c} + (x_{1,c-1} + \lambda_1\sigma_{21} - \mu_1\tau^\mu_{1,c})\tau^\lambda_{1,c} + ... \\ &(x_{1,c-1} + \lambda_1\sigma_{21} + \tfrac{1}{2}\lambda_1(\sigma_{12} + \tau^\mu_{2,c} + \tau^\lambda_{2,c}) - ... \\ &\mu_1\tau^\mu_{1,c})(\sigma_{12} + \tau^\mu_{2,c} + \tau^\lambda_{2,c}), \quad c = 1, ..., C, \qquad (13a) \\ w_{2,c} =&(x_{2,c-1} + \tfrac{1}{2}\lambda_2(\sigma_{21} + \tau^\mu_{1,c} + \tau^\lambda_{1,c} + \sigma_{12}))(\sigma_{21} + ... \\ &\tau^\mu_{1,c} + \tau^\lambda_{1,c} + \sigma_{12}) + (x_{2,c-1} + \lambda_2(\sigma_{21} + \tau^\mu_{1,c} + ... \\ &\tau^\lambda_{1,c} + \sigma_{12}) - \tfrac{1}{2}\mu_2\tau^\mu_{2,c})\tau^\mu_{2,c} + (x_{2,c-1} + \lambda_2(\sigma_{21} + ... \\ &\tau^\mu_{1,c} + \tau^\lambda_{1,c} + \sigma_{12}) - \mu_2\tau^\mu_{2,c})\tau^\lambda_{2,c}, \quad c = 1, ..., C. \qquad (13b) \end{aligned}$$

Using (13), the transient costs (10), considering C cycles, can be written as

$$J_p(C) = C(s_{12} + s_{21}) + Q_p(C).$$

Here, $Q_p(C)$ is the solution to the quadratic programming problem, for C cycles, given by

$$Q_p(C) = \min_{\tau^\mu_{i,c}, \tau^\lambda_{i,c}} \sum_{i=1}^{2}\sum_{c=1}^{C} c_i^+ w_{i,c} - J_s^*(\sigma_{ij} + \tau^\mu_{i,c} + \tau^\lambda_{i,c}), i \neq j, \tag{14}$$

subject to constraints (11a)-(11b) and

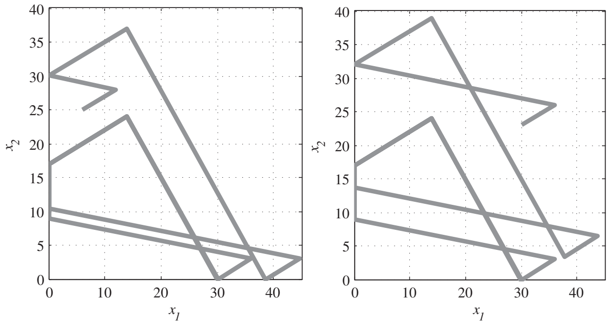
$$\begin{aligned} x_{1,c} &\geq \lambda_1(\sigma_{12} + \tau^\mu_{2,c}, \tau^\lambda_{2,c}), & c = 1, ..., C, \qquad (15a) \\ x_{2,c} &\geq 0, & c = 1, ..., C, \qquad (15b) \\ x_{1,c-1} &\leq x_1^{\max} - \lambda_1\sigma_{21}, & c = 1, ..., C, \qquad (15c) \\ x_{2,c} &\leq x_2^{\max} - (\mu_2 - \lambda_1)\tau^\mu_{2,c}, & c = 1, ..., C, \qquad (15d) \end{aligned}$$

where constraints (15a)-(15b) follow from $x_i(t) \geq 0$ and constraints (15c)-(15d) follow from (11c).
To negate the termination effect, we calculate the costs of the first $C - 2$ cycles of the optimal trajectory for C cycles, resulting from (14), and denote these costs by $J_p^*(C)$. In other words, the transient trajectory for C cycles is derived, and the costs of the first $C - 2$ cycles only are considered.

Given a system with the initial state outside the steady-state trajectory, the number of cycles required to derive the optimal transient trajectory is not easily determined, as shown by, for instance, the optimal trajectory in Figure 8. However, a lower bound on the number of cycles required for a feasible transient trajectory $C^{\min}$ can be determined by using a clearing policy, regarding the initial state and considering a system without capacity or service time constraints. For a system with capacity or service time constraints, this number of cycles is usually not enough to reach the steady-state trajectory. Starting from this lower bound, and by adding extra cycles, we solve the QP problem until a feasible transient trajectory is derived. Note that this transient trajectory is not necessarily the optimal trajectory, i.e., adding more cycles can lower the costs. Therefore, the number of cycles considered in the QP problem (14) is increased until the total costs required for the transient trajectory to reach the steady-state trajectory does no longer change, i.e., $J_p^*(C) = J_p^*(C + i)$, $\forall i > 0$. Then, adding more cycles does not result in a different transient trajectory, in the sense that it only adds steady-state cycles to the solution. Hence, we can then conclude that the transient solution is the optimal one.

For the system with parameters (9), and without constraints on queue length, cycle time and service times, the optimal transient trajectory for initial state $[1, 6, 25]$ is presented in Figure 7a. Note, that the initial mode is setting up to serve queue 1, and that the steady-state trajectory is reached during the second cycle. It can be seen that for this initial state a clearing policy (until the steady-state trajectory is reached) yields the optimal performance. However, the optimal trajectory for initial state $[1, 30, 23]$, presented in Figure 7b, gives a different result. First, after the setup, queue 1 is emptied. Second, after the setup, queue 2 is served until a content of 3.43 is reached, then the system switches to serve queue 1. Note that this queue is not emptied. Next, queues 1 and 2 are both cleared before reaching the steady-state trajectory.

For the trajectory depicted in Figure 7b, it is clearly shown that a trade-off exists between a build-up of the much more expensive queue 1 and switching before emptying queue 2. This behavior is not present in symmetric systems, as a clearing policy is optimal for symmetric systems, see for instance [1, 9].



(a) Initial state $[1, 6, 25]$. (b) Initial state $[1, 30, 23]$.

Figure 7: Optimal transient trajectories with different initial states. In (a) the clearing policy is optimal, in (b) it is not.

Each optimal transient trajectory contains *switching points*. A switching point is the state $[m, x_1, x_2]$ at which the system switches to serve the other queue, i.e., switching between modes $m = 2$ and $m = 3$ and between modes $m = 4$ and $m = 1$. Experimentally combining the switching points of optimal trajectories, i.e., solving the transient problem for a set of initial states and collecting the switching points, results in a *switching curve*. A switching curve characterizes the optimal transient structure for any given initial state. Note that for a system with service time constraints, i.e., with at least one of the constraints (11a)-(11b), switching curves may not exist in general, as the switching points are affected by the constraints and will depend on the initial state. For the system without service time constraints, switching curves can possibly be derived analytically as follows. Starting from the steady-state trajectory, an area can be characterized from which the transient trajectory converges with a single operation to the steady-state trajectory. Next, an area can be characterized for which the system converges to the steady-state trajectory in two steps, and the optimal service times can be derived. Continuing these steps might result in the switching curves.

The (experimentally determined) switching curves for the system with parameters (9) are presented in Figure 8, along with a trajectory with initial state $[1, 40, 80]$. The switching curve for a transition between modes $m = 2$ and $m = 3$ is given by the line $x_1 = 0$ and $x_2 \geq 17$, where $(0, 17)$ is the switching point of the optimal steady-state trajectory, presented by point D in Figure 2. The switching curve for a transition between modes $m = 4$ and $m = 1$ is discontinuous with linear segments. These segments do not overlap, i.e., each initial state has a single optimal trajectory.
For the system with parameters (9) and $c_1^+ = 4$, the switching curve is continuous, see Figure 12. Here, the switching curve for a transition between modes $m = 4$ and $m = 1$ is piecewise linear.

Adding maximal queue length constraints $x_1^{\max} = 75$ and $x_1^{\max} = 92$ to this model results in the switching curves depicted in Figure 10. The figure also displays the optimal transient trajectory with initial state $[1, 40, 80]$. It can be seen that the switching curves, originating from the queue level constraints, are located $\lambda_i \sigma_{ji}$ below $x_i^{\max}$, as the queue length increases during the setup. Note that the initial queue contents, for starting in mode 1, are limited to $x_1(0) \leq x_1^{\max} - \lambda_1 \sigma_{21}$ and $x_2(0) \leq x_2^{\max} - \lambda_2(\sigma_{21} + \sigma_{12})$.

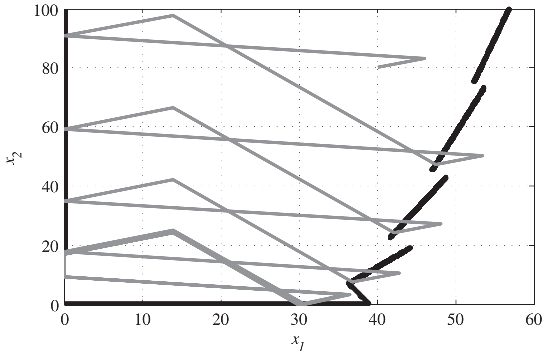


Figure 8: Discontinuous switching curves (black), for the system with parameters (9), and transient trajectory (gray) for initial state $[1, 40, 80]$.

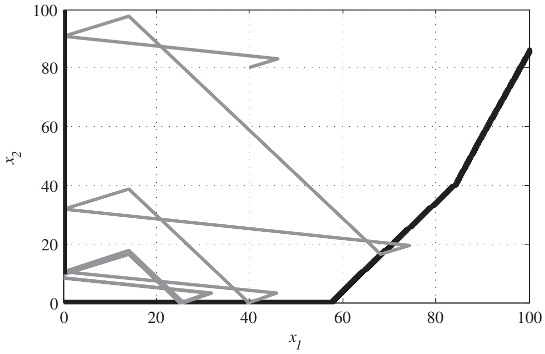


Figure 9: Continuous switching curves (black), for the system with parameters (9) and $c_1^+ = 2$, and transient trajectory (gray) with initial state $[1, 40, 80]$.

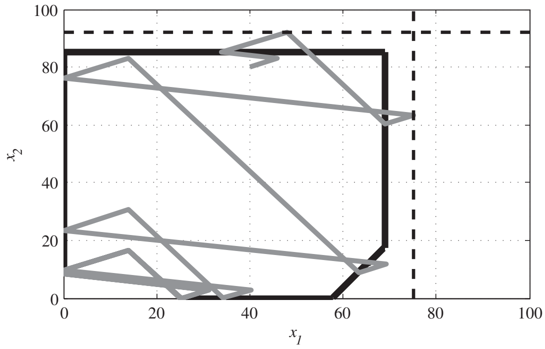


Figure 10: Switching curves (black) for the system with parameters (9), $c_1^+ = 2$, $x_1^{\max} = 75$ and $x_1^{\max} = 92$ with transient trajectory (gray) for initial state $[1, 40, 80]$.

For an optimal transient policy, the switching curves can be used to indicate the switching moments. From our experiments we find that for $c_i^+ \mu_i \geq c_j^+ \mu_j$, queue i is always emptied and the optimal policy for $c_i^+ \mu_i = c_j^+ \mu_j$ is, as expected, a clearing policy (unless prohibited by restrictions (11a)-(11c)).

Furthermore, for an optimal transient policy, also the optimal initial mode (given contents $x(0)$), if it is not predefined, can be derived. Together with the switching curves, this gives the policy for optimal transient behavior given initial queue contents. Once the optimal initial mode is known, the queues are served until a switching point is reached, switch to the successive mode, until converging to the optimal steady-state trajectory. If all initial modes are allowed, the setup modes $m(0) = 1$ and $m(0) = 3$ are of course not optimal. Therefore, a comparison of the transient costs starting with both modes $m(0) = 2$ and $m(0) = 4$ results in the optimal initial mode. For the system with parameters (9), the optimal initial modes are presented in Figure 11, along with the switching curves. For initial queue contents in the gray area the optimal initial mode is $m(0) = 2$, $m(0) = 4$ otherwise.

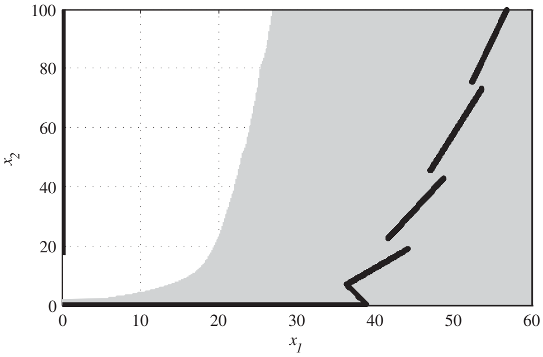


Figure 11: Switching curves (black) and optimal initial mode for the system with parameters (9), $m(0) = 2$ in the gray area, $m(0) = 4$ otherwise.

Also for the system with parameters (9), where $c_1^+ = 2$ and queue length constraints $x_1^{\max} = 75$ and $x_1^{\max} = 92$, the optimal initial modes are presented in Figure 12. The dark gray area indicates that $m(0) = 2$ is optimal and the light gray area indicates optimal mode $m(0) = 4$. For the remaining area, no trajectories exist as these would violate the queue constraints.

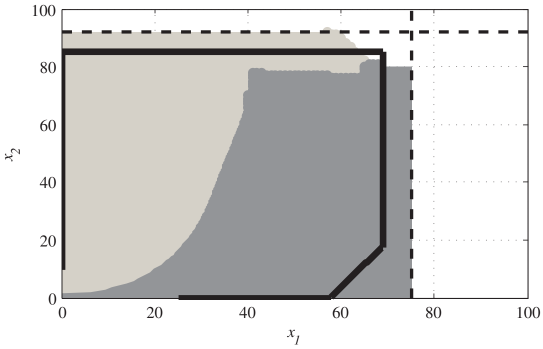


Figure 12: Switching curves (black) and optimal initial mode for the system with parameters (9), $c_1^+ = 2$, $x_1^{\max} = 75$ and $x_1^{\max} = 92$. $m(0) = 2$ in the dark gray area and $m(0) = 4$ in the light gray area.

5. CONCLUSIONS

In this paper we studied the optimal steady-state and transient trajectories for a two queue switching server. The steady-state problem is formulated as a QP, given a fixed cycle time. By solving the QP over a range of cycle times, the optimal steady-state trajectory is derived. An advantage of this method is that it is very flexible in adding objectives and constraints, e.g., setup times and/or setup costs, including backlog and constraints on cycle times, service times and queue lengths.

Second, we formulated the transient problem, i.e., a transient trajectory which minimizes the costs of reaching the optimal steady-state trajectory. This is also formulated as a QP, depending on the number of cycles to calculate. Evaluating over a range of cycles results in the optimal transient trajectory given initial queue contents and initial mode. For the system without capacity constraints, switching curves can be derived by combining switching points of optimal trajectories, i.e., points at which the system switches to serve other queues. These switching curves are the blueprint of a policy for optimal transient behavior. Furthermore, the optimal initial mode, if not predefined, can be derived. Together with the switching curves, this gives the policy for optimal transient behavior.

This work can be extended to systems with more than two queues. For a fixed queue routing, i.e., fixed service order of queues, the approach can be easily extended by adding the extra queues. If the queue routing is not fixed, multiple routes are possible to converge to the steady-state trajectory. For all these routes, the optimal service periods can be derived using our approach. Then, the best result yields the optimal behavior. Also, analytical derivation of the switching curves is a suggestion for further study on this topic.

6. ACKNOWLEDGMENTS

This work is supported by the Netherlands Organization for Scientific Research (NWO-VIDI grant 639.072.072).

7. REFERENCES

[1] M. Boccadoro and P. Valigi. A switched system model for the optimal control of two symmetric competing queues with finite capacity. In L. Menini, L. Zaccarian, and C. Abdallah, editors, *Current Trends in Nonlinear Systems and Control*, Systems and Control: Foundations and Applications, pages 455–474. Birkhauser Boston, 2006.

[2] M. Del Gaudio, F. Martinelli, and P. Valigi. A scheduling problem for two competing queues with finite capacity and non-negligible setup times. In *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, volume 3, pages 2355–2360 vol.3, 2001.

[3] J. Haddad, B. De Schutter, D. Mahalel, I. Ioslovich, and P.-O. Gutman. Optimal Steady-State Control for Isolated Traffic Intersections. *Automatic Control, IEEE Transactions on*, 55(11):2612–2617, 2010.

[4] J. Haddad, P.-O. Gutman, I. Ioslovich, and D. Mahalel. Discrete dynamic optimization of N-stages control for isolated signalized intersections. *Control Engineering Practice*, 21(11):1553 – 1563, 2013.

[5] M. Hofri and K. Ross. On the optimal control of two queues with server setup times and its analysis. *SIAM Journal on Computing*, 16(2):399–420, 1987.

[6] V. Imbastari, F. Martinelli, and P. Valigi. An optimal scheduling problem for a system with finite buffers and non-negligible setup times and costs. In *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, volume 1, pages 1156– 1161 vol.1, dec. 2002.

[7] J. Kimemia and S. B. Gershwin. An algorithm for the computer control of a flexible manufacturing system. *IIE Transactions*, 15(4):353–362, 1983.

[8] Z. Liu, P. Nain, and D. Towsley. On optimal polling policies. *Queueing Systems*, 11:59–83, 1992.

[9] F. Martinelli and P. Valigi. Dynamic scheduling for a single machine system under different setup and buffer capacity scenarios. *Asina Journal of Control*, 6(2):229–241, jun 2004.

[10] J. van Eekelen. *Modelling and control of discrete event manufacturing flow lines*. PhD thesis, Eindhoven, University of Technology, 2008.

[11] J. van Eekelen, E. Lefeber, and J. Rooda. Feedback control of 2-product server with setups and bounded buffers. In *Proceedings of the 2006 American Control Conference*, pages 544–549, 2006.

[12] D. van Zwieten, E. Lefeber, and I. Adan. Optimal periodical behavior of a multiclass fluid flow network. In *Proceedings of the Conference on Management and Control of Production and Logistics*, 2013.

[13] Y. Yang, B. Mao, S. Chen, S. Liu, and M. Liu. Effect of bus rapid transit signal priority effect on traffic flow. *Shenzhen University Science and Engineering*, 30(1):91–97, 2013.