OPTIMAL BEHAVIOR FOR THE KUMAR-SEIDMAN NETWORK OF SWITCHING SERVERS

Erjen Lefeber
Department of Mechanical Engineering
Eindhoven University of Technology
The Netherlands
A.A.J.Lefeber@tue.nl

J.E. Rooda
Department of Mechanical Engineering
Eindhoven University of Technology
The Netherlands
J.E.Rooda@tue.nl

Abstract

In this paper we consider the two server switching network introduced by Kumar and Seidman. We consider the problem of minimizing the weighted average wip in the system, assuming non-increasing costs downstream. Assuming that both servers have the same period, optimal network behavior has been derived. By means of an illustrative example is shown that this optimal network behavior at first glance can be counter-intuitive. In particular this implies that currently existing ways for controlling these kind of networks do not achieve optimal network behavior.

Key words

Control of networks, Deterministic multi-class fluid queues, Hybrid dynamical systems, Optimal control, Setup times, Switched servers.

1 Introduction

Consider a network of servers through which different types of jobs flow. One could think of a manufacturing system, i.e. a network of machines through which different types of products flow. An other example would be an urban road network of crossings with traffic lights through which cars flow. A third example would be a network of computers through which different streams of data flow.

These networks might show some unexpected behavior. In (Banks and Dai, 1997) was shown by simulation that even when each server has enough capacity, these networks can be unstable in the sense that the wip in the network explodes as time evolves. Whether this happens or not depends on the policy used to control the flows through the network. In (Kumar and Seidman, 1990) was shown analytically that using a clearing policy (serve the queue you are currently serving until it is empty, then switch to another queue) certain networks become unstable, even for deterministic systems with no setup times.

In (Perkins and Kumar, 1989) several clearing policies have been introduced, the so-called Clear a Fraction (CAF) policies. It was shown that these policies are stable for a single server in isolation in a deterministic environment. Furthermore, it was shown that a CAF policy stabilizes a multi server system, provided the network is acyclic. A network is called acyclic if the servers can be ordered in such a way that wip can only move from one server to a server higher in the ordering. A network is called non-acyclic if such an ordering is not possible. The example in (Kumar and Seidman, 1990) shows that non-acyclic networks exist that cannot be stabilized by a CAF policy.

The main reason why CAF policies can fail for a non-acyclic network is because they spend too long on serving one type. This results in starvation of other servers and therefore a waste of their capacity. Due to this waste the effective capacity of these other servers is not sufficient anymore, resulting in an unstable system. This observation has led to the development of so-called buffer regulators (Humes, 1994; Perkins *et al.*, 1994) or gated policies. The main idea is that each buffer contains a gate, so the buffer is split into two parts (before and after the gate). Instead of switching depending on the total buffer contents, switching is now determined based on the buffer contents after the gate. As a result, a server might now leave a buffer earlier, avoiding long periods of serving one type. It has been shown in (Perkins *et al.*, 1994) that under certain conditions on these regulators the (possibly non-acyclic) network is stabilized. Since non-acyclic networks are only unstable under certain conditions, applying buffer regulators is not always necessary. Needlessly applying buffer regulators results in a larger mean wip in the network, which from a performance point of view is undesired. Furthermore, it is not known whether these policies result in optimal network behavior.

In (Savkin, 1998) a different approach has been developed. First the minimal period is determined during which the network is able to serve all wip that ar-

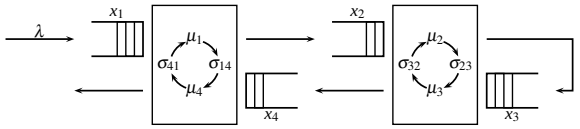

Figure 1. The system introduced in (Kumar and Seidman, 1990).

rives during that period. This minimal period then determines how much time to spend on each type, resulting in a time table which determines when each server should be serving which type. It was shown in (Matveev and Savkin, 2000; Savkin, 1998) that if each server follows this time schedule (possibly idling if no wip of the scheduled type is available), the system behavior becomes regular. In particular this implies that the system converges towards a periodic orbit. It was not yet known if optimal network behavior could be achieved. The example in Section 4 illustrates that this is not always the case.

In (Eekelen *et al.*, 2006) we considered the most simple network of switching servers: a single server which serves only two types. Starting from the goal to minimize the time-averaged weighted wip in the system, we derived optimal periodic network behavior. Furthermore, a controller was presented which made the system converge towards this optimal periodic behavior. Determining optimal system behavior for more than two types already is challenging problem, see also (Takagi, 1986; Takagi, 1990; Takagi, 1997). In this paper we extend the results from (Eekelen *et al.*, 2006) to the network of two servers as introduced by Kumar and Seidman in (Kumar and Seidman, 1990).

This paper is organized as follows. In Section 2 the Kumar-Seidman network as introduced in (Kumar and Seidman, 1990) is presented, as well as a way of modeling this network by means of a hybrid dynamical control system with constraints. In Section 3 optimal periodic network behavior is derived. For a particular choice of parameters in Section 4 the results from Section 3 are applied. This example illustrates that currently existing policies for controlling these kinds of networks do not achieve optimal network behavior. Finally, Section 5 concludes this paper.

2 The Kumar-Seidman network

Consider the manufacturing system shown in Figure 1. A single type is considered which first visits server 1, then server 2, then server 2 again, and finally server 1 again. The successive buffers visited will be denoted by 1, 2, 3, and 4, respectively. A constant input rate $\lambda > 0$ into buffer 1 is assumed, while the maximal processing rates at the buffers are $\mu_1 > 0$, $\mu_2 > 0$, $\mu_3 > 0$, and $\mu_4 > 0$, respectively. For ease of exposition we also introduce $\rho_i = \lambda/\mu_i$ ($i \in \{1,2,3,4\}$). Lastly, the times for setting-up buffers 1 and 4 at server 1 are $\sigma_{41} > 0$ and $\sigma_{14} > 0$, the times for setting-up to buffers 2 and 3 at server 2 are $\sigma_{32} > 0$ and $\sigma_{23} > 0$. Even when for this system each server has enough capacity, i.e. both $\rho_1 + \rho_4 < 1$ and $\rho_2 + \rho_3 < 1$, it has been shown in (Kumar and Seidman, 1990) that whenever $\rho_2 + \rho_4 > 1$, using a clearing policy for both servers yields an unstable system.

Assumption 1. *Throughout the remainder of this paper we restrict ourselves to this situation, i.e. we assume that*

$$\frac{\mu_2\mu_4}{\mu_2+\mu_4} < \lambda < \min\left(\frac{\mu_1\mu_4}{\mu_1+\mu_4}, \frac{\mu_2\mu_3}{\mu_2+\mu_3}\right). \quad (1)$$

We model the network by means of a hybrid fluid model. The state of this system is not only given by the buffer contents $x_i \in \mathbb{R}$ ($i \in \{1,2,3,4\}$), but also by the remaining setup time at server j, $x_0^j \in \mathbb{R}$ ($j \in \{1,2\}$), and the current mode $m = (m^1, m^2) \in \{(1,2),(1,3),(4,2),(4,3)\}$. We say that the system is in mode (1,2) when server 1 is (being) set-up for step 1 and server 2 is (being) set-up for step 2. Similarly for the other modes.

The input of this system is given by rates $u_1 \le \mu_1$, $u_2 \le \mu_2$, $u_3 \le \mu_3$, and $u_4 \le \mu_4$, at which respectively buffers 1, 2, 3, and 4 are being served (a server not necessarily has to serve at full rate), as well as the current activity for server 1, $u_0^1 \in \{❶, ①, ❹, ④\}$, and for server 2, $u_0^2 \in \{❷, ②, ❸, ③\}$. The activity ❶ denotes a setup for serving step 1, whereas ① denotes serving step 1. Similarly the activities for step 2, 3, and 4 can be distinguished.

As mentioned above, the dynamics of this system is hybrid. On the one hand we have the discrete event dynamics

$$\begin{aligned}
&x_0^1 := \sigma_{14}; \quad m^1 := 4 \quad \text{if } u_0^1 = ❹ \text{ and } m^1 = 1\\
&x_0^1 := \sigma_{41}; \quad m^1 := 1 \quad \text{if } u_0^1 = ❶ \text{ and } m^1 = 4\\
&x_0^2 := \sigma_{23}; \quad m^2 := 3 \quad \text{if } u_0^2 = ❸ \text{ and } m^2 = 2\\
&x_0^2 := \sigma_{32}; \quad m^2 := 2 \quad \text{if } u_0^2 = ❷ \text{ and } m^2 = 3.
\end{aligned}$$

In words: if the system is currently in a mode, and according to the input the current activity becomes "setup to a different mode", both the remaining setup time and current mode change.

On the other hand we have the continuous dynamics

$$\dot{x}_0^1(t) = \begin{cases} -1 & \text{if } u_0^1 \in \{❶, ❹\} \\ 0 & \text{if } u_0^1 \in \{①, ④\} \end{cases} \qquad \dot{x}_0^2(t) = \begin{cases} -1 & \text{if } u_0^2 \in \{❷, ❸\} \\ 0 & \text{if } u_0^2 \in \{②, ③\} \end{cases}$$

$$\begin{aligned}
&\dot{x}_1(t) = \lambda - u_1(t) && \dot{x}_2(t) = u_1(t) - u_2(t)\\
&\dot{x}_4(t) = u_3(t) - u_4(t) && \dot{x}_3(t) = u_2(t) - u_3(t).
\end{aligned}$$

Furthermore, at each time instant the input is subject

to the constraints $u_1 \geq 0$, $u_2 \geq 0$, $u_3 \geq 0$, $u_4 \geq 0$, and

$$
\begin{array}{llll}
u_0^1 \in \{❶,❹\} & u_1 = 0 & u_4 = 0 & \text{for } x_0^1 > 0 \\
u_0^1 \in \{①,❹\} & u_1 \leq \mu_1 & u_4 = 0 & \text{for } x_0^1 = 0, x_1 > 0, m^1 = 1 \\
u_0^1 \in \{①,❹\} & u_1 \leq \lambda & u_4 = 0 & \text{for } x_0^1 = 0, x_1 = 0, m^1 = 1 \\
u_0^1 \in \{❶,④\} & u_1 = 0 & u_4 \leq \mu_1 & \text{for } x_0^1 = 0, x_4 > 0, m^1 = 4 \\
u_0^1 \in \{❶,④\} & u_1 = 0 & u_4 \leq u_3 & \text{for } x_0^1 = 0, x_4 = 0, m^1 = 4 \\
u_0^2 \in \{❷,❸\} & u_2 = 0 & u_3 = 0 & \text{for } x_0^2 > 0 \\
u_0^2 \in \{②,❸\} & u_2 \leq \mu_1 & u_3 = 0 & \text{for } x_0^2 = 0, x_2 > 0, m^2 = 2 \\
u_0^2 \in \{②,❸\} & u_2 \leq u_1 & u_3 = 0 & \text{for } x_0^2 = 0, x_2 = 0, m^2 = 2 \\
u_0^2 \in \{❷,③\} & u_2 = 0 & u_3 \leq \mu_3 & \text{for } x_0^2 = 0, x_3 > 0, m^2 = 3 \\
u_0^2 \in \{❷,③\} & u_2 = 0 & u_3 \leq u_2 & \text{for } x_0^2 = 0, x_3 = 0, m^2 = 3.
\end{array}
$$

In words, these constraints say that in case the server is setting-up, no wip can be served. Furthermore, in case a setup has been completed, only the step can be processed for which the server has been set-up. This processing takes place at a rate which is at most μ_i if wip of step i is available in the buffer and at most atthe arrival rate if no wip of step i are available in the buffer ($i \in \{1,2,3,4\}$). Also, it is possible to either stay in the current mode, or to switch to the other mode. In particular it is possible during setup to leave that setup and start a setup to the other step again. The latter setup is assumed to take the entire setup time.

3 Optimal network behavior

Having defined the state, input, dynamics and constraints for the system, we can consider the problem of deriving optimal behavior for this system. To that end, we consider the goal of minimizing

$$J = \limsup_{t\to\infty} \frac{1}{t} \int_0^t c_1 x_1(\tau) + c_2 x_2(\tau) + c_3 x_3(\tau) + c_4 x_4(\tau)\, \mathrm{d}\tau \tag{2}$$

with $c_1 \geq c_2 \geq c_3 \geq c_4 > 0$. That is, we consider the problem of minimizing the time-averaged weighted wip in the system with the restriction that downstream wip is not weighted more heavily than upstream wip.

Under this assumption we can derive the following lemmas.

Lemma 2. *Without loss of generality it can be assumed that servers always serve at the highest possible rate, after which they might idle, i.e. process wip at rate zero. This highest possible rate equals μ_i when the buffer contains wip ($x_i > 0$), or the arrival rate to that server (which might be zero, but not necessarily) otherwise.*

Proof. Suppose that a policy is given for which after having completed the setup to step i, buffer i contains a wip of x_i^0 and at the end of serving step i, buffer i contains a wip of x_i^f. Then one can consider the alternative policy which serves step i equally long and first serves at the highest possible rate, i.e. at the maximal processing rate as long as the buffer contains wip or at the arrival rate in case the buffer is empty. In the end, this alternative policy idles to make sure that at the end of serving step i the buffer contains a wip of x_i^f. Clearly, while serving step i at rate μ_i the wip in the buffer cannot decrease faster (or increase slower in case the server feeding into step i currently serves at a higher rate than μ_i) and in the end cannot increase faster than in this alternative strategy. Therefore, for the alternative policy at each time instant the wip for step i is minimal. In particular, if the given policy is different, the wip for step i is less at certain time instants. Since the time evolution of the other steps remains the same for both policies and serving wip does not increase costs, costs cannot be higher using the alternative strategy.

Lemma 3. *Without loss of generality it can be assumed that servers never idle at the end of serving step i.*

Proof. Suppose that a server would idle at the end of serving step i. After serving step i it switches to serving step $5-i$. Furthermore, assume that this server stops serving step $5-i$ at time t_f. Consider an alternative policy which does not idle at the end of serving step i, but switches immediately to serving step $5-i$ and stays in this mode until time t_f, serving an equal amount of wip as the supposed optimal strategy. For this alternative strategy the evolution of x_i does not change. Also $x_{5-i}(t_f)$ is equal. However, (some of) the wip of step $5-i$ might be served sooner. Therefore costs cannot be higher for the alternative strategy.

Corollary 4. *Without loss of generality it can be assumed that servers only idle when the buffer of the currently served step is empty and no wip of that step is arriving.*

Assumption 5. *Throughout the remainder of this paper we restrict ourselves to periodic behavior where each server serves its both steps exactly once. In particular this implies that minimizing* (2) *reduces to minimizing*

$$J = \frac{1}{T} \int_0^T c_1 x_1(\tau) + c_2 x_2(\tau) + c_3 x_3(\tau) + c_4 x_4(\tau)\, \mathrm{d}\tau \tag{3}$$

where $c_1 \geq c_2 \geq c_3 \geq c_4 > 0$ and T denotes the period of this periodic behavior, satisfying

$$T \geq \max\left(\frac{\sigma_{14} + \sigma_{41}}{1-\rho_1-\rho_4}, \frac{\sigma_{23} + \sigma_{32}}{1-\rho_2-\rho_3} \right)$$

to guarantee existence of periodic behavior.

Lemma 6. *Without loss of generality it can be assumed that server 1 successively goes through the following actions*

- ① *at rate μ_1, for a duration of $\tau_1^\mu = \frac{\rho_1}{1-\rho_1}[\rho_4 T + \sigma_{14} + \sigma_{41}]$,*

- ① *at rate* λ *for a duration of* $\tau_1^\lambda = \frac{1}{1-\rho_1}[(1-\rho_1-\rho_4)T-(\sigma_{14}+\sigma_{41})]$,
- ❹ *for a duration of* σ_{14},
- ④ *at rate* μ_4 *for a duration of* $\tau_4 = \rho_4 T$,
- ❶ *for a duration of* σ_{41}.

Proof. From (1) it follows that $\mu_4 < \mu_3$ as the function $x \mapsto \mu_2 x/(\mu_2+x)$ is strictly increasing for $x > 0$. From lemmas 2 and 3 and it then follows that server 1 can serve step 1 only first at rate μ_1 and then at rate λ, whereas step 4 can only be served first at rate 0 (only when $x_4 = 0$) and then at rate μ_4.

Instead of serving step 4 at rate 0 as long as $x_4 = 0$, server 1 might as well continue serving step 1 longer for this amount of time, moving wip from server 1 to server 2 which does not increase costs.

The durations of the actions can be determined from the requirements that each step needs to serve the wip that arrives during the period, and total service and setups cover the entire period:

$$\lambda T = \mu_1 \tau_1^\mu + \lambda \tau_1^\lambda \tag{4a}$$
$$\lambda T = \mu_4 \tau_4 \tag{4b}$$
$$T = \tau_1^\mu + \tau_1^\lambda + \sigma_{14} + \tau_4 + \sigma_{41}. \tag{4c}$$

Lemma 7. *Without loss of generality it can be assumed that server 2 successively goes through the following actions*

- ② *at rate* 0, *for a duration of* $\tau_2^0 = (1-\rho_2-\rho_3)T-(\sigma_{23}+\sigma_{32})$,
- ② *at rate* μ_2 *for a duration of* $\tau_2^\mu = \rho_2 T$,
- ❸ *for a duration of* σ_{23},
- ③ *at rate* μ_3 *for a duration of* $\tau_3 = \rho_3 T$,
- ❷ *for a duration of* σ_{32}.

Proof. From (1) it follows that $\mu_2 < \mu_1$ as the function $x \mapsto \mu_4 x/(\mu_4+x)$ is strictly increasing for $x > 0$. From lemmas 2 and 3 and it then follows that server 2 can serve step 2 at rate 0, at rate μ_2 and at rate λ, whereas step 3 can only be served first at rate μ_3.

Let the successive total durations of service be denoted by τ_2^0, τ_2^μ, τ_2^λ and τ_3. From the requirements that each step needs to serve the wip that arrives during the period, and total service and setups cover the entire period, we obtain:

$$\tau_2^\mu = \frac{\rho_2}{1-\rho_2}[\rho_3 T + \tau_2^0 + \sigma_{23} + \sigma_{32}] \tag{5a}$$
$$\tau_2^\lambda = \frac{1}{1-\rho_2}[(1-\rho_2-\rho_3)T - \tau_2^0 - (\sigma_{23}+\sigma_{32})] \tag{5b}$$
$$\tau_3 = \rho_3 T. \tag{5c}$$

Assume that $\tau_2^\lambda > 0$. The only way that server 2 can produce at rate λ, is when also server 1 produces at rate λ. Before server 2 can serve at rate λ it first needs to clear buffer x_2. This observation results in the requirement that

$$\lambda(\tau_1^\mu + \tau_1^\lambda - \frac{\mu_1}{\mu_2}\tau_1^\mu - \tau_2^\lambda) = \mu_2(\tau_2^\mu - \frac{\mu_1}{\mu_2}\tau_1^\mu).$$

Substituting (4) and (5) results in

$$\frac{[\mu_2-\mu_1][\mu_4(\sigma_{14}+\sigma_{41})+\lambda T]\lambda^2}{\mu_2\mu_4(\mu_1-\lambda)} = 0,$$

which has no feasible solutions. Therefore, $\tau_2^\lambda = 0$.

The durations of the actions readily follow from the requirements that each step needs to serve the wip that arrives during the period, and total service, idling, and setups cover the entire period.

Lemma 8. *For optimal periodic behavior given a period* $T > \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ *we have:*

$$\int_0^T x_1(\tau)\,\mathrm{d}\tau = \frac{\lambda}{2(1-\rho_1)}(\rho_4 T + \sigma_{14} + \sigma_{41})^2 \tag{6a}$$
$$\int_0^T x_2(\tau)\,\mathrm{d}\tau = \frac{1}{2}\lambda(\rho_2-\rho_1)T^2 - \frac{1}{2}\lambda(1-\rho_1)\tau_1^{\lambda\,2} \tag{6b}$$
$$\int_0^T x_3(\tau)\,\mathrm{d}\tau = \frac{1}{2}(\rho_2+\rho_3)\lambda T + \sigma_{23}\lambda T \tag{6c}$$
$$\int_0^T x_4(\tau)\,\mathrm{d}\tau = (\mu_4-\lambda)\tau^{43}T + \frac{1}{2}\lambda(\rho_4-\rho_3)T^2, \tag{6d}$$

where τ_1^λ *is as given in Lemma 6 and*

$$\tau^{43} = (\rho_2+\rho_4-1)T + \sigma_{23} + \sigma_{41}$$

denotes the amount of time that service of step 4 is started earlier than service of step 3.

Proof. When server 1 completes serving step 1, $x_1 = 0$. For a duration of $\rho_4 T + \sigma_{14} + \sigma_{41}$ step 1 is not being served, resulting in an increase to $\lambda(\rho_4 T + \sigma_{14} + \sigma_{41})$, which then decreases to 0 again during τ_1^μ. This results in (6a).

By assumption $\tau_2^0 > 0$, i.e. server 2 idles. From Lemma 3 we know that this can only be when $x_2 = 0$ and server 2 waits for server 1 to start serving step 1. Furthermore, since $\rho_2+\rho_4 > 1$, we have that $\tau_1^\mu + \tau_1^\lambda + \sigma_{14} < \rho_2 T$, i.e. server 1 already starts serving step 4 before server 2 completes serving step 2. In particular this implies that x_2 increases from 0 to $(\mu_1-\mu_2)\tau_1^\mu$ for a duration of τ_1^μ. Next, it decreases from $(\mu_1-\mu_2)\tau_1^\mu$ to $(\mu_1-\mu_2)\tau_1^\mu + (\lambda-\mu_2)\tau_1^\lambda$ for a duration of τ_1^λ. Finally, it decreases from $(\mu_1-\mu_2)\tau_1^\mu + (\lambda-\mu_2)\tau_1^\lambda$ to 0 again for a duration of $\rho_2 T - \tau_1^\mu - \tau_1^\lambda$. This results in (6b).

When server 2 completes serving step 3, $x_3 = 0$. For a duration of $\sigma_{32} + \tau_2^0$ nothing happens. Next, during

$\rho_2 T$ the buffer contents x_3 increase to a value of λT. For a duration of σ_{23} we have $x_3 = \lambda T$, after which during $\rho_3 T$ the buffer contents decrease to 0 again. This results in (6c).

Since service of step 4 is started earlier than service of step 3, the initial buffer contents of buffer 4 should be such that $x_4 = 0$ at the moment service of step 3 starts, since x_4 starts to increase from that moment on as $\mu_3 > \mu_4$. Now two cases can be considered. Either $\sigma_{32} + \tau_2^0 \leq \sigma_{41}$ or $\sigma_{32} + \tau_2^0 \geq \sigma_{41}$.

First, consider the case $\sigma_{32} + \tau_2^0 \leq \sigma_{41}$. Then we have that x_4 decreases from $\mu_4 \tau^{43}$ to 0 for a duration of τ^{43}. Next, it increases from 0 to $(\mu_3 - \mu_4)[(1-\rho_2)T - \sigma_{23} - \sigma_{41}]$ for a duration of $(1-\rho_2)T - \sigma_{23} - \sigma_{41}$, followed by a further increase from $(\mu_3 - \mu_4)[(1-\rho_2)T - \sigma_{23} - \sigma_{41}]$ to $\mu_4 \tau^{43}$ for a duration of $\sigma_{23} + \sigma_{41} - (1-\rho_2 - \rho_3)T$. Finally, $x_4 = \mu_4 \tau^{43}$ for a duration of $(2 - \rho_2 - \rho_3 - \rho_4)T - \sigma_{41} - \sigma_{23}$.

Second, consider the case $\sigma_{32} + \tau_2^0 \geq \sigma_{41}$. Then we also have that x_4 decreases from $\mu_4 \tau^{43}$ to 0 for a duration of τ^{43}. But next, it increases from 0 to $(\mu_3 - \mu_4)\rho_3 T$ for a duration of $\rho_3 T$, followed by a decrease from $(\mu_3 - \mu_4)\rho_3 T$ to $\mu_4 \tau^{43}$ for a duration of $(1 - \rho_2 - \rho_3)T - \sigma_{23} - \sigma_{41}$. Finally, $x_4 = \mu_4 \tau^{43}$ for a duration of $(1-\rho_4)T$.

Both alternatives result in (6d).

For period $T = \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ we have $\tau_2^0 = 0$. Therefore, not necessarily server 2 starts serving step 2 at exactly the time at which server 1 starts serving step 1. Let t denote the amount of time that server 1 starts serving step 1 later than server 2 starts serving step 2.

Lemma 9. *For optimal periodic behavior given a period $T = \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ and $0 \leq t \leq T$ we have* (6a), (6c) *and:*

$$\bar{x}_2 = \begin{cases} x_2^0 + (\mu_2 - \lambda)Tt & \text{if } 0 \leq t \leq \rho_2 T \\ x_2^0 + \lambda T(T-t) & \text{if } \rho_2 T \leq t \leq T \end{cases} \tag{7a}$$

$$\bar{x}_4 = \begin{cases} x_4^0 - (\mu_4 - \lambda)Tt & \text{if } 0 \leq t \leq \tau^{43} \\ x_4^0 - \mu_4 T(\tau^{43} - \rho_4 t) & \text{if } \tau^{43} \leq t \leq \tau^{43} + (1-\rho_4)T \\ x_4^0 + (\mu_4 - \lambda)T(T-t) & \text{if } \tau^{43} + (1-\rho_4)T \leq t \leq T, \end{cases} \tag{7b}$$

where $\bar{x}_i$ is an abbreviation for $\int_0^T x_i(\tau)d\tau$ ($i \in \{2,3\}$) and

$$\bar{x}_2 = \frac{1}{2}\lambda(\rho_2 - \rho_1)T^2 - \frac{1}{2}\lambda(1-\rho_1)\tau_1^{\lambda\,2}$$
$$\bar{x}_4 = (\mu_4 - \lambda)\tau^{43}T + \frac{1}{2}\lambda(\rho_4 - \rho_3)T^2$$

i.e. the expressions (6b) *and* (6d).

Proof. Similar to the proof of the previous lemma.

Now we have all ingredients for determining optimal periodic behavior for the system as described in Section 2. We more or less can start from the results from lemmas 8 and 9 and optimize over all possible values for T (and t).

First we restrict ourselves to the case

$$\frac{\sigma_{14} + \sigma_{41}}{1 - \rho_1 - \rho_4} > \frac{\sigma_{23} + \sigma_{32}}{1 - \rho_2 - \rho_3}.$$

Then we have $T > \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$, so we can restrict ourselves to the results from Lemma 8. From this lemma we know that

$$\frac{1}{T}\int_0^T x_1(\tau)\,\mathrm{d}\tau = \alpha_{1,2}T + \alpha_{1,1} + \alpha_{1,0}\frac{1}{T} \tag{8a}$$
$$\frac{1}{T}\int_0^T x_2(\tau)\,\mathrm{d}\tau = \alpha_{2,2}T + \alpha_{2,1} - \alpha_{2,0}\frac{1}{T} \tag{8b}$$
$$\frac{1}{T}\int_0^T x_3(\tau)\,\mathrm{d}\tau = \alpha_{3,2}T + \alpha_{3,1} \tag{8c}$$
$$\frac{1}{T}\int_0^T x_4(\tau)\,\mathrm{d}\tau = \alpha_{4,2}T + \alpha_{4,1}, \tag{8d}$$

where

$$\alpha_{1,2} = \frac{\lambda\rho_4^2}{2(1-\rho_1)} \qquad \alpha_{2,2} = \frac{1}{2}\lambda(\rho_2 - \rho_1)$$
$$\alpha_{1,1} = \frac{\lambda\rho_4(\sigma_{14}+\sigma_{41})}{(1-\rho_1)} \qquad \alpha_{2,1} = \frac{\lambda(1-\rho_1-\rho_4)(\sigma_{41}+\sigma_{14})}{(1-\rho_1)}$$
$$\alpha_{1,0} = \frac{\lambda(\sigma_{14}+\sigma_{41})^2}{2(1-\rho_1)} \qquad \alpha_{2,0} = \frac{\lambda(\sigma_{14}+\sigma_{41})^2}{2(1-\rho_1)}$$
$$\alpha_{3,2} = \frac{1}{2}\lambda(\rho_2 + \rho_3) \qquad \alpha_{4,2} = (\mu_4 - \lambda)(\sigma_{41} + \sigma_{23})$$
$$\alpha_{3,1} = \lambda\sigma_{23} \qquad \alpha_{4,1} = (\mu_4 - \lambda)(\rho_2 + \rho_4 - 1) + \frac{1}{2}\lambda(\rho_4 - \rho_3)$$

Notice that all $\alpha_{i,j} > 0$, and that $\alpha_{1,0} = \alpha_{2,0}$. This implies that (8b), (8c) and (8d) are strictly increasing functions of T. In particular we have that if $c_1 = c_2$, (3) is minimized for $T = \frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4}$. In case $c_1 > c_2$ we need to determine a local minimum for the function $\alpha_2 T + \alpha_1 + \alpha_0/T$ where

$$\alpha_2 = c_1\alpha_{1,2} + c_2\alpha_{2,2} + c_3\alpha_{3,2} + c_4\alpha_{4,2}$$
$$\alpha_1 = c_1\alpha_{1,1} + c_2\alpha_{2,1} + c_3\alpha_{3,1} + c_4\alpha_{4,1}$$
$$\alpha_0 = (c_1 - c_2)\alpha_{1,0}$$

This minimum is achieved for $T = \sqrt{\alpha_0/\alpha_2}$.

The above derivations can be summarized in the following

Proposition 10. *Consider the system as described in Section 2, satisfying assumptions 1 and 5. Furthermore, assume that $\frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4} > \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$. Then the period T of the periodic orbit minimizing* (3) *is equal to*

- $\frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4}$ *when either* $c_1 = c_2$ *or* $\sqrt{\alpha_0/\alpha_2} \leq \frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4}$

- $\sqrt{\alpha_0/\alpha_2}$ *when both* $c_1 > c_2$ *and* $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4}$.

where α_0 *and* α_2 *are given by the above equations.*

Furthermore, the periodic orbit starts serving step 1 and step 2 at full rate simultaneously, and the durations of the consecutive modes are as described in lemmas 6 and 7.

Next, we consider the case

$$\frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4} \leq \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}.$$

Similar to the derivation of Proposition 10 we have that the period T of the periodic orbit minimizing (3) is equal to $\frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ when either $c_1 = c_2$ or $\sqrt{\alpha_0/\alpha_2} \leq \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$. However, when both $c_1 > c_2$ and $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ the period T of the periodic orbit minimizing (3) is not necessarily equal to $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$. It is when $\mu_4 \leq \mu_2 + \frac{c_2-c_4}{c_4}(\mu_2-\lambda)$, however in case $\mu_4 > \mu_2 + \frac{c_2-c_4}{c_4}(\mu_2-\lambda)$ an other possibility exists. From (7) it can be seen that in the latter case $c_2\bar{x}_2 + c_4\bar{x}_4$ is a decreasing function of t for $0 \leq t \leq \tau^{43}$. Using a period of $T = \sqrt{\alpha_0/\alpha_2}$ results in

$$J = 2\sqrt{\alpha_0\alpha_2} + \alpha_1. \tag{9}$$

On the other hand, using a period of $T = \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ with $t = \tau^{43}$ results in

$$J = \alpha_2 T + \alpha_1 + \frac{\alpha_0}{T} + [c_2(\mu_2-\lambda) - c_4(\mu_4-\lambda)]T\tau^{43}. \tag{10}$$

Depending on whether (9) or (10) results in the smallest value, the optimal period can be determined.

Proposition 11. *Consider the system as described in Section 2, satisfying assumptions 1 and 5. Furthermore, assume that* $\frac{\sigma_{14}+\sigma_{41}}{1-\rho_1-\rho_4} \leq \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$. *Then the period* T *of the periodic orbit minimizing* (3) *is equal to*

- $\frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ *when either* $c_1 = c_2$ *or* $\sqrt{\alpha_0/\alpha_2} \leq \frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$
- $\sqrt{\alpha_0/\alpha_2}$ *when both* $c_1 > c_2$, $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{32}+\sigma_{23}}{1-\rho_2-\rho_3}$, *and* $\mu_4 \leq \mu_2 + \frac{c_2-c_4}{c_4}(\mu_2-\lambda)$
- $\sqrt{\alpha_0/\alpha_2}$ *when both* $c_1 > c_2$, $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{32}+\sigma_{23}}{1-\rho_2-\rho_3}$, $\mu_4 > \mu_2 + \frac{c_2-c_4}{c_4}(\mu_2-\lambda)$ *and* (9) *is greater than* (10)
- $\frac{\sigma_{23}+\sigma_{32}}{1-\rho_2-\rho_3}$ *when both* $c_1 > c_2$, $\sqrt{\alpha_0/\alpha_2} > \frac{\sigma_{32}+\sigma_{23}}{1-\rho_2-\rho_3}$, $\mu_4 > \mu_2 + \frac{c_2-c_4}{c_4}(\mu_2-\lambda)$ *and* (9) *is less than* (10)

where α_0 *and* α_2 *are given by the above equations and* τ^{43} *as defined in Lemma 8.*

Furthermore, in the first three cases the periodic orbit starts serving step 1 and step 2 at full rate simultaneously, whereas in the fourth case the periodic orbit starts serving step 4 and step 3 at full rate simultaneously. The durations of the consecutive modes are as described in lemmas 6 and 7.

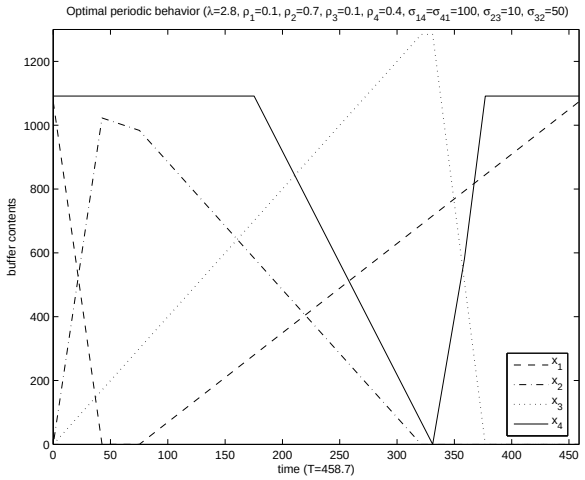

Figure 2. Optimal periodic behavior for $\lambda = 2.8$, $\rho_1 = 0.1$, $\rho_2 = 0.7, \rho_3 = 0.1, \rho 4 = 0.4, \sigma_{14} = \sigma_{41} = 100, \sigma_{23} = 10$, $\sigma_{32} = 50, c_1 = 10000, c_2 = 3, c_3 = 2, c_4 = 1$.

4 Example

In the previous section we derived optimal network behavior for the case presented in Section 2. In this section we make a possible choice for the parameters and show the corresponding optimal network behavior (under assumptions 1 and 5).

Consider the case where $\lambda = 2.8$, $\rho_1 = 0.1$, $\rho_2 = 0.7$, $\rho_3 = 0.1$, $\rho 4 = 0.4$, $\sigma_{14} = \sigma_{41} = 100$, $\sigma_{23} = 10$, $\sigma_{32} = 50$. For the cost function we assume that $c_1 = 10000$, $c_2 = 3$, $c_3 = 2$, $c_4 = 1$. The resulting optimal periodic behavior is given in Figure 2. In this figure we see that from 0 till 42 both step 1 and step 2 are served at maximal rate (μ_1 and μ_2 respectively). Since server 1 serves at a higher rate than server 2 we see not only a decrease of x_1 and an increase of x_3, but also an increase of x_2. At $t = 42$ buffer 1 becomes empty and server 1 continues serving step 1, but now at the arrival rate. As a result, x_2 starts to decrease. At $t = 75$, server 1 stops serving step 1 and starts a setup to step 4. As a result, x_2 decreases even faster. At $t = 175$, server 1 has completed its setup and starts serving step 4, causing x_4 to decrease. At $t = 321$, buffer 2 becomes empty and server 2 switches to serving step 3. Service of step 3 starts at $t = 331$, exactly at the time that buffer 4 runs empty. Since step 3 is served at a higher rate than step 4, buffer 4 increases even though server 1 is still serving step 4. At $t = 358$, server 1 stops serving step 4 and start its setup to step 1. As server 2 is still serving step 3, the buffer contents of x_4 start to increase at an even higher rate. At $t = 377$, buffer 3 becomes empty and server 2 starts a setup to step 2 which is completed at $t = 427$. From $t = 427$ until $t = 458$, server 2 idles. Machine 1 completes its setups at $t = 458$, after which

the whole cycle starts all over again.

One of the important observations to make is that both servers seem to be wasting capacity. Machine 1 is serving step 1 at the arrival rate from $t = 42$ till $t = 75$. Machine 2 idles from $t = 427$ till $t = 458$. At first glance this seems rather strange for optimal periodic behavior. How can it be optimal to waste capacity at both servers? A first observation is that the minimal process cycle of server 1 would be 400 times units, whereas the minimal process cycle of server 2 would be 300 time units. Therefore it is not surprising that at server 2 capacity is wasted. But why is capacity wasted at server 1? Actually two ways exist of wasting capacity that need to be considered. One way of wasting capacity is by serving at a less than maximal rate. But an other way of wasting capacity is by having a short period. In the latter case on the average more time is wasted on setups. Given a total setup time per cycle of 200 per period, for a period of 400 time units server 1 spends 50% of its time on setups. Whereas for a period of 800 time units, only 25% of the time is spend on setups. So on the one hand one can waste capacity by serving at a lower rate, on the other hand capacity can be wasted by setting-up most of the time. Apparently a trade-off exists, which in this case results in a period of $T = 458$.

5 Conclusions

In this paper we considered optimal network behavior for the hybrid system introduced in (Kumar and Seidman, 1990). After introducing the system and describing its dynamics, we considered the problem of minimizing the weighted average wip in the system, assuming non-increasing costs downstream. Assuming that both servers have the same period, optimal network behavior has been derived. By means of an illustrative example it was shown that this optimal network behavior at first glance can be counterintuitive. In particular this implies that currently existing ways for controlling these kind of networks do not achieve optimal network behavior. An next step will be to derive controllers that make the network converge towards this optimal network behavior. A possible approach to this problem has been introduced in (Lefeber and Rooda, 2006), and worked out for the system under consideration in this paper only for a specific choice of parameters in (Lefeber and Rooda, 2008). This approach generally leads to non-distributed network controllers. That is, knowledge of the global network state is required to control all servers simultaneously. It is a challenge to derive distributed controllers that make the network converge to a priori specified behavior. For the specific choice of parameters considered in (Lefeber and Rooda, 2008) such a distributed controller can be determined. Extending this to a more general setting would be the subject of further research.

Acknowledgements

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

References

Banks, J. and J.G. Dai (1997). Simulation studies of multiclass queueing networks. *IIE Transactions* **29**, 213–219.

Eekelen, J.A.W.M. van, E. Lefeber and J.E. Rooda (2006). Feedback control of 2-product server with setups and bounded buffers. In: *Proceedings of the American Control Conference*. Minneapolis, Minnesota, USA.

Humes, Jr, C. (1994). A regulator stabilization technique: Kumar Seidman revisited. *IEEE Transactions on Automatic Control* **39**(1), 191–196.

Kumar, P.R. and T.I. Seidman (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* **35**(3), 289–298.

Lefeber, E. and J.E. Rooda (2006). Controller design of switched linear systems with setups. *Physica A*.

Lefeber, E. and J.E. Rooda (2008). Controller design for flow networks of switched servers with setup times: The Kumar-Seidman case as an illustrative example. *Asian Journal of Control*.

Matveev, A.S. and A.V. Savkin (2000). *Qualitative theory of hybrid dynamical systems*. Control Engineering. Birkhäuser. Boston, Massachusetts, USA.

Perkins, J. and P.R. Kumar (1989). Stable, distributed, real-time scheduling of flexible manufacturing/assembly/disassembly systems. *IEEE Transactions on Automatic Control* **34**(2), 139–148.

Perkins, J.R., C. Humes, Jr and P.R. Kumar (1994). Distributed scheduling of flexible manufacturing systems: Stabiliy and performance. *IEEE Transactions on Robotics and Automation* **10**(2), 133–141.

Savkin, A.V. (1998). Regularizability of complex switched server queueing networks modelled as hybrid dynamical systems. *Systems and Control Letters* **35**, 291–299.

Takagi, H. (1986). *Analysis of Polling Systems*. MIT Press. Cambridge, Massachusetts, USA.

Takagi, H. (1990). Queueing analysis of polling models: An update. In: *Stochastic Analysis of Computer and Communication Systems* (H. Takagi, Ed.). Chap. 1, pp. 267–318. Elsevier Science Publishers B.V.. Amsterdam, The Netherlands.

Takagi, H. (1997). Queueing analysis of polling models: Progress in 1990–1994. In: *Frontier in Queueing: Models and Applications in Science and Engineering* (J.H. Dshalalow, Ed.). Chap. 5, pp. 119–146. CRC Press. New York, USA.