SUBOPTIMAL SWITCHING CONTROL OF QUEUING SYSTEMS

J. Smieja

Dept. of Automatic Control, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland
e-mail: jsmieja@ia.polsl.gliwice.pl

Abstract: This paper is concerned with a problem of control in the queuing systems. Rather than deal with mean waiting time or average queue length, which is the most often used approach to this problem, it concentrates on transient states and minimizing probability of long queue. First a model of a queuing system with controlled service intensity is analyzed. Subsequently, a system with multiple service stations that can be switched in or off is introduced. For both systems, the optimal control problem is formulated in L^1 space and necessary conditions for optimal control are presented.

Keywords: Queuing systems, optimal control, bang-bang control..

1. INTRODUCTION

Despite long history of research (see eg. (Gross and Harris, 1998; Medhi, 2003) and references therein) and rich literature devoted to problems of modeling of queuing systems, most of it concentrates either on theoretical aspects which applicability is arguable, or models based on analysis of mean probabilities or average length queue (Sennott, 1999; Athuraliya *et al*., 2001, Hollot *et al.*, 2002). However, due to rapid development of applications, especially in computer networks or telecommunication systems, analysis of transient states becomes increasingly important. In most of the literature, however, transient states are analyzed only in context of most standard models or in context of average length queue or waiting time. Moreover, the results published so far address mainly linearized models, while this paper deals with nonlinear ones.

Although the main applications of the work seem to be in the fields of computer network or telecommunication systems analysis and design, the models are referred to as queuing systems, to underscore the theoretical aspect of this paper. Moreover, for the same reason, instead of using *Active Queue Management* term, standard for TCP networks modeling, the concept of general control is used, which highlights the original field of control theory that is applied for analysis.

2. SERVICE RATE CONTROL.

Let us suppose that in the analyzed queuing system service rate can be controlled. It can be achieved, for example by directing incoming requests to stations that have different performance (with cost of use or leasing them increasing with growing efficiency). The system description can be presented in the following way:

$$\begin{cases} \dot{P}_0(t) = -\lambda_0 P_0(t) + [1+u(t)]\mu_1 P_1(t) \\ \dot{P}_1(t) = \lambda_0 P_0(t) - (\lambda_1 + [1+u(t)]\mu_1) P_1(t) + \\ \qquad + [1+u(t)]\mu_2 P_2(t) \\ \dots \\ \dot{P}_i(t) = \lambda_{i-1} P_{i-1}(t) - (\lambda_i + [1+u(t)]\mu_i) P_i(t) \\ \qquad + [1+u(t)]\mu_{i+1} P_{i+1}(t) \\ \qquad\qquad\qquad\qquad \text{for } 1 \le i < l-1 \\ \dot{P}_{l-1}(t) = \lambda_{l-2} P_{l-2}(t) - (\lambda_{l-1} + [1+u(t)]\mu_{l-1}) P_{l-1}(t) \\ \qquad + \mu P_l(t) \\ \dot{P}_i(t) = \lambda P_{i-1}(t) - (\lambda+\mu) P_i(t) + \mu P_{i+1}(t) \\ \qquad\qquad\qquad\qquad \text{for } i \ge l \\ \dots \end{cases} \tag{1}$$

where $P_i(t)$ denotes the probability that the length of the queue at the time instant t is equal to i, λ, λ_i, μ and μ_i are model parameters, $0 \le u(t) \le u_{\max}$ represents control effect on the service intensity. It seems justifiable to assume that the difference in the efficiency will be visible only for small length of the queue. Though usually service and arrival rates are assumed to be independent of queue length, here they can vary for the first l equations to underscore wide applicability of the method presented in the paper.

Though similar examples could be found e.g. in (Sennott, 1999), In this paper a unique performance index will be introduced to evaluate quality of control system. The aim is to minimize probability of a queue longer than arbitrarily chosen length l, simultaneously taking into account the cumulative cost of the control, i.e.

$$\min \leftarrow J = \sum_{i \ge l} P_i(T) + r\int_0^T u(t)dt \qquad (2)$$

Arbitrary parameter l determines decomposition illustrated in the Fig.1.

Such infinite dimensional system description is very similar to the models analyzed in biomedical modeling (Swierniak *et al.*, 2003) that are based on branching processes. Applying methodology developed for dealing with those models it is possible to both find the transient states for this system and solve the optimal control problem.

To make analysis of such models possible it is convenient to present it in the form of a block diagram shown in Fig.1, effectively decomposing the model into two parts The first one, of finite dimension, does not require parameters to meet any particular assumptions. The second subsystem is infinite dimensional, with tridiagonal system matrix, and does not contain terms containing control $u(t)$.

Using methods similar to that shown in our previous works devoted to biomedical modeling (Swierniak *et al.*, 1998, 1999), it is possible to derive the following transfer function in the model (1):

$$K_1(s) = \frac{P_l(s)}{P_{l-1}(s)} = \frac{\lambda}{\mu}\frac{s-(\lambda+\mu)-\sqrt{[s-(\lambda+\mu)]^2-4\lambda\mu}}{2\lambda} \qquad (3)$$

Moreover, if the initial length of the queue is k, i.e. initial condition is given by. $P_k(0) = 1$, $P_i(0) = 0$ for $i \ne k$,

$$\sum_{i \ge l} P_i(t) = P_\Sigma^k(t) + P^+(t) \qquad (4)$$

where $P_\Sigma^k(t)$ is free component, equal to 0 if $k < l$, otherwise defined by

$$P_\Sigma^k(t) = P_k(0)\cdot \left[1-(k-l+1)\left(\sqrt{\frac{\mu}{\lambda}}\right)^{k-l+1}\int_0^t \frac{I_{k-l+1}\left(2\sqrt{\mu\lambda}\,\tau\right)}{\tau}\exp[-(\lambda+\mu)\tau]d\tau\right] \qquad (5)$$

$I_k(.)$ denotes modified Bessel function of the k–th order.

$P^+(t)$ is a forced component given by

$$P^+(t) = \mu\int_0^t k_\Sigma^1(t-\tau)P_{l-1}(\tau)d\tau \qquad (6)$$

where k_Σ is given by the same relation as $P_\Sigma^k(t)$ for $k = l$ with $P_l(0)$ set to one.

The system description (1) in the form of infinite number of ODEs is not very convenient for the optimization procedure. Instead, it can be transformed into integro-differential one.

Let us denote

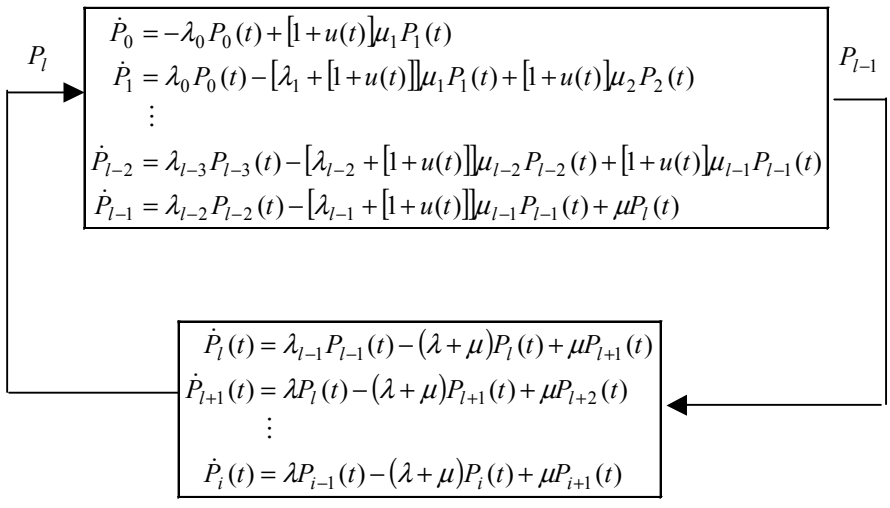

Figure 1. Decomposition of the general system model

$$\widetilde{P} = \begin{bmatrix} P_0 \\ \vdots \\ P_{l-1} \end{bmatrix} \tag{7}$$

Let us also assume the initial conditions $P_i(0) = 0$ for $i > l-1$ (the assumption is justified in queuing systems). Then, the last equation in the first subsystem, influenced directly by control, as presented on Fig. 1, can be transformed into an integro-differential form:

$$\begin{aligned} \dot{P}_{l-1}(t) = \mu_{l-1} P_{l-1}(t) u_{l-1}(t) + \mu P_l(t) + \\ + a_{l-1} c_1 \int_0^t k_1(t-\tau) P_{l-1}(\tau) d\tau \end{aligned} \tag{8}$$

where $k_1(t)$ is the inverse Laplace transform of $K_1(s)$, given by

$$k_1(t) = \frac{1}{\mu}\left(\sqrt{\frac{\mu}{\lambda}}\right) \frac{I_1\left(2\sqrt{\mu\lambda}\, t\right)}{t} \exp\left[-(\lambda+\mu)t\right] \tag{9}$$

Similarly, other equations can also be rewritten in the same way leading to the transformation of the model (1) into the following form:

$$\dot{\widetilde{P}} = h(u, \widetilde{P}) + \int_0^t \widetilde{f}(t-\tau) P_{l-1}(\tau) d\tau\,, \tag{10}$$

where $h(..), \widetilde{f}(.)$ - vector functions

$$h_0(u, \widetilde{P}) = -\lambda_0 P_0(t) + \mu_1 u(t) P_1(t) \tag{11}$$

$$\begin{aligned} h_k(u, \widetilde{P}) = \lambda_{k-1} P_{k-1}(t) - \left[\lambda_k + \mu_k u(t)\right] P_k(t) + \\ + \mu_{k+1} u(t) P_{k+1}(t) \end{aligned} \tag{12}$$

for $1 \le k < l-1$

$$h_{l-1}(u, \widetilde{P}) = \lambda_{l-2} P_{l-2}(t) - \left[\lambda_{l-1} + \mu_{l-1} u(t)\right] P_{l-1}(t) \tag{13}$$

$$\widetilde{f}_k(t-\tau) = \begin{cases} 0 & for \quad k < l-1 \\ \lambda_l k_1(t-\tau) & for \quad k = l-1 \end{cases} \tag{14}$$

for $k = 0,1,\ldots, l-1$.

It is important to notice that, although the performance index (2) seems to consist of two components - a sum and an integral, the sum actually involves another integral that stems from (4)–(5). Therefore, it should be rewritten to emphasize this relation. Substituting (4) and (5) into (2) we obtain:

$$J = P_\Sigma^1(T) + \int_0^T \left[P_\Sigma^1(T-\tau) P_{l-1}(\tau) + ru(\tau)\right] d\tau \tag{15}$$

A number of formulations of necessary conditions for the optimization problem for dynamical systems governed by integro-differential equations can be found in literature (e.g. Gabasov and Kirilova, 1971; Curtain and Zwart, 1995). However, they usually either are too general to be efficiently applied in such particular problem or have too strong constraints for example smoothness of the control function. Nevertheless, following the line of reasoning presented in (Bate, 1969), it is possible to derive the necessary conditions for optimal control:

$$\begin{aligned} u^{opt}(t) = \arg\min_u \Big[ru(t) + \\ + p^T(t) h(u, \widetilde{P}) + \lambda_l \int_t^T p_{l-1}(\tau) k_1(t-\tau) P_{l-1}(\tau) d\tau \Big] \end{aligned} \tag{16}$$

$$\begin{aligned} \dot{p}(t) = -\Big[q^T(t) + p^T(t) h_{\widetilde{P}}(\widetilde{P}(t), u(t)) + \\ + \int_t^T p^T(\tau) \widetilde{f}(t-\tau) d\tau \Big], \end{aligned} \tag{17}$$

$$q(t) = \begin{bmatrix} 0 & \ldots & 0 & \lambda P_\Sigma^l(T-t) \end{bmatrix}^T$$

$$p_i(T) = 1,\ i = 0,1,\ldots,\ l-1 \tag{18}$$

$p(t)$ – adjoint vector.

Taking into account the control constraint and bilinear form of (11)–(13), it can be proved that, in order to satisfy (16), the optimal control must be bang-bang one. Then, to find optimal number of switches and switching times, a gradient method can be developed, following the line of reasoning presented in (Smieja *et al*. 2000).

3. CONTROLLED *M/M/C/K* SYSTEM

The standard description of the *M/M/c/K* queuing system is given by (Kleinrock, 1976)

$$\begin{cases} \dot{P}_0(t) = -\lambda P_0(t) + \mu_i P_1(t) \\ \dot{P}_i(t) = \lambda P_{i-1}(t) - (\lambda + \mu_i) P_i(t) + \mu_{i+1} P_{i+1}(t) \\ \qquad\qquad for\ 1 \le i < K \end{cases} \tag{19}$$

where

$$\mu_i = \begin{cases} i\mu & for \ \ i \le c \\ c\mu & for \ \ i \ge c \end{cases} \tag{20}$$

The control in this case can be interpreted as switching on and off appropriate service lines. For the sake of simplicity, let us assume that there are only two of them available, i.e. $u(t) = \{1,2\}$. The problem of optimal control can be defined as minimizing the performance index (2) for the system described by

$$\begin{cases} \dot{P}_0(t) = -\lambda P_0(t) + \mu P_1(t) \\ \dot{P}_i(t) = \lambda P_{i-1}(t) - (\lambda + \mu u(t)) P_i(t) + \mu u(t) P_{i+1}(t) \\ \qquad\qquad for\ 1 \le i \le K \end{cases} \tag{21}$$

To find the conditions that optimal control should satisfy and subsequently apply them to find a solution, it is convenient to assume that $u(t) \in [1,2]$.

Then, in order to derive necessary conditions standard maximum principle can be applied (Pontryagin,). The Hamiltonian is given by

$$H = ru(t) - p_0(t)\lambda P_0(t) + \sum_{i=1}^{K-1}\left[p_i(t)\left(\lambda P_{i-1}(t) - \left[\lambda + \mu u(t)\right]P_i(t) + \mu u(t)P_{i+1}(t)\right)\right] - p_K(t)\left[\lambda P_{K-1}(t) + \mu u(t)P_K(t)\right] \quad (22)$$

where p denotes the adjoint vector. The conjugate equation takes the following form

$$\begin{cases} \dot{p}_0(t) = \lambda\left(p_0(t) - p_1(t)\right) \\ \dot{p}_i(t) = -\mu u(t)p_{i-1}(t) + \left[\lambda + \mu u(t)\right]p_i(t) \\ \qquad\qquad - \lambda p_{i+1}(t) \qquad\qquad 0 < i < K \\ \dot{p}_K(t) = -\mu u(t)p_K(t) \end{cases} \quad (23)$$

$$p_i(T) = \begin{cases} 0 & for \quad i < l \\ 1 & for \quad l-1 \le i \le K \end{cases} \quad (24)$$

Since the control u is bounded and Hamiltonian is linear with respect to it, the optimal control must take a *bang-bang* form. Actually, if one goes back to the initial problem formulation, such solution is the only admissible. However, proceeding in the way shown here, one can develop a gradient method for finding the optimal solution.

The case with more service stations is much more complex, since the control involved can take values from a set of more than two elements. Therefore, in addition to finding optimal number of switches and switching times, the decision must be made about how many stations should be switched on or off. In that case, the optimization problem must be reformulated as the searching for optimal switches between different system structures, while the general system description for the model with n available service lines would be given by:

$$\begin{cases} \dot{P}_0(t) = -\lambda P_0(t) + \mu P_1(t) \\ \dot{P}_i(t) = \lambda P_{i-1}(t) - (\lambda + \mu \min\{i, u(t)\})P_i(t) + \\ \qquad + \mu \min\{i+1, u(t)\}P_{i+1}(t) \\ \qquad\qquad\qquad\qquad for\ 1 \le i \le K \\ \dot{P}_K(t) = \lambda P_{K-1}(t) - (\lambda + \mu \min\{i, u(t)\})P_K(t) \end{cases} \quad (25)$$

u(t)∈ {1,2,…,n}

However, the solution to the optimization problem stated above is not yet available. Nevertheless, since control variable has to take integer values, once again it can be tested numerically.

4. THE GRADIENT METHOD

Let u^* denote the bang-bang control ($u^*=u$ for the model given in the section 2, while for the model from section 3 $u^*=u-1$)

Let τ_i denote the i-th switching time. Then the control u^* can be presented as

$$u^*(t) = \sum_{i=0}^{M/2}\left[\mathbf{1}(t-\tau_{2i}) - \mathbf{1}(t-\tau_{2i+1})\right], \quad (26)$$

where $\mathbf{1}(.)$ is a unit step function, M denotes the number of switches. Even and odd subscripts correspond to switching on and off, correspondingly (therefore M can be assumed to be an even number, since it is always possible to force switching off at the end of time horizon).

Let $\tilde{H}$ denote Hamiltonian given by (22) in the section 3 or the argument of arg min in (16). It is crucial to notice that

$$\Delta J \approx \int_0^T \frac{\partial \tilde{H}}{\partial u}\delta u dt \quad (27)$$

From (Duda, 1995) it follows that:

$$\Delta J \approx 2\sum_{j=1}^{M}(-1)^{j+1}\left.\frac{\partial \tilde{H}}{\partial u}\right|_{\tau=\tau_j}\delta\tau_j \quad (28)$$

where $\delta\tau_i$ denotes the variation of the switching time τ_i.

Hence, to minimize the performance index J

$$\delta\tau_j = (-1)^j k_j \left.\frac{\partial \tilde{H}}{\partial u}\right|_{\tau=\tau_j} \quad (29)$$

$j = 1,2, \ldots, M$, k_j - positive number

Taking this into account the following algorithm can be applied:

1. Assume number of switches, as well as initial switching times.

2. Solve the equation describing the system dynamics for bang-bang control with assumed switching times.

3. Compute $p(t)$ from the adjoint equation (25) integrating it backward in time.

4. Calculate values of $\delta\tau_i$ from (29).

5. Compute new switching times $\tau_i + \delta\tau_i$.

6. If two switching times are close to each other, reject them and modify appropriately the number of switches.

7. Repeat steps 2-5 until stop condition is satisfied, e.g. $\sum_{j=0}^{M}(\delta\tau_j)^2 < \varepsilon$, where ε is small given number.

In both problems, it is assumed that there are no singular solutions. Since their non-existence is not proven here, obtained results can be regarded only as suboptimal (optimal in the set of admissible switching controls). One must remember, however,

that in the case of infinite dimensional model and bang-bang control, finding the optimal number of switches is almost impossible. Even for finite dimensional problem, given as the second problem, the sufficient conditions are not known to the author. Nevertheless, starting from arbitrary large number of switches (justified by model application) the suboptimal results should be satisfactory.

6. CONCLUSIONS

In this paper we are concerned with two different models of queuing systems. Explicit control variable has been introduced to both of them, making it possible to define and solve optimization problem with the performance index defined in l^1 space of summable sequences. In both cases, the resulting solution is in the form of the open-loop bang-bang control. A gradient method has been proposed to find optimal switching times for both models.

Contrary to the approach widely used in the literature, addressing average queue lengths or waiting times, the method proposed in the paper deals directly with state variables.

In the first, infinite-dimensional model, control has been used to change the service rate. It was built upon the general M/M/1 model, with the underlying assumption that the service rate can be increased only if the length of the queue had not crossed som threshold level. For that system, basing on model decomposition, it was possible to transform its description into integro-differential form. That, in turn, allows solving an optimal control problem The methodology presented in the paper makes it possible to address more general problem, in which control is multidimensional, i.e. the the rate of service can be independently changed for different queue lengths. Using exactly the same way of reasoning as presented in this paper, it is possible to derive necessary conditions for optimal control in that case.

In the second, finite-dimensional model, control consists in switching on/off additional service stations. The case of two service stations has been addressed, while more general model has been proposed, to be analyzed in the future.

ACKNOWLEDGEMENTS

This work has been partially supported by KBN grant No 3 T11A 029 28 in 2005.

REFERENCES

Athuraliya S., V.H. Li, S.H. Low and Q. Yin (2001). REM:Active queue management, *IEEE Network Mag.*, **15**, 48-53.

Bate R.B. (1969). The optimal control of systems with Transport Lag. *Advances in Control Systems*. **7**, 165-224.

Curtain R.F. and H.J. Zwart (1995). *An introduction to infinite-dimensional linear systems theory*. Springer Verlag, New York.

Duda Z. (1995). A gradient method for application of chemotherapy protocols. *Journal of Biological Systems*. **3**(1), 3–11.

Gabasov R. and F.M. Kirilowa (1971). *Qualitative Theory of Optimal Processes.* Moscow: Nauka.

Gross D. and C.M. Harris (1998). *Fundamentals of queuing theory*. New York: Wiley.

Hollot C.V., V. Misra, D.Towsley and W. Gong (2002). Analysis and Design of Controllers for AQM Routers Supporting TCP Flows. *IEEE Trans. on Automatic Control*. **47**(6), 945-959.

Kleinrock L. (1976). *Queuing Systems.Vol.1: Theory*. New York: Wiley.

Mehdi J. (2003). *Stochastic models in queuing theory*. Amsterdam; Boston: Academic Press.

Sennott L.I. (1999). *Stochastic dynamic programming and the control of queuing systems*. New York: Wiley.

Smieja J., A. Swierniak and Z. Duda (2000). Gradient Method for Finding Optimal Scheduling in Infinite Dimensional Models of Chemotherapy, *Journal of Theoretical Medicine*. **3**, 25-36.

Swierniak A., M. Kimmel and A. Polanski (1998). Infinite dimensional model of evolution of drug resistance of cancer cells, *Journal of Mathematical Systems, Estimation and Control*, **8**(1), 1–17.

Swierniak A., Polanski A., Kimmel M., Bobrowski A. and Smieja J. (1999): *Qualitative analysis of controlled drug resistance model - inverse Laplace and semigroup approach*, Control and Cybernetics, **28**, 61-74.

Swierniak A., A. Polanski, J. Smieja and M. Kimmel (2003). Modelling growth of drug resistant cancer populations as the system with positive feedback, *Mathematical and Computer Modelling*, **37**(11), 1245-1252