# Analysis of admission control mechanisms using non-linear control theory

Kihl, Maria; Robertsson, Anders; Wittenmark, Björn

*Published in:*
Proceedings of the 8th IEEE International Symposium on Computers and Communication. Vol. 2

*DOI:*
10.1109/ISCC.2003.1214294

2003

# Analysis of Admission Control Mechanisms using Non-linear Control Theory

Maria Kihl[1], Anders Robertsson[2] and Björn Wittenmark[2]

[1]*Department of Communication Systems,* [2]*Department of Automatic Control*

*Lund Institute of Technology, BOX 118, 221 00 Lund, Sweden*

maria@telecom.lth.se, {andersro, bjorn}@control.lth.se

## Abstract

*All service control nodes can be modelled as a server system with one or more servers processing incoming requests. In this paper we show how non-linear control theory may be used when analyzing admission control mechanisms for server systems. Two models are developed, one linear and one non-linear. We show that, due to the non-linearities appearing in a real server system, linear control theory is insufficient when designing controllers for these systems. With non-linear analysis, however, the dynamics of a server system may be analysed and taken care of by choosing the controller parameters appropriately.*

## 1. Introduction

Service control nodes, that is nodes that contain service logic and control, play an important role in most modern communication networks. From the application layer view point, a service control node consists of a server system with one or more servers processing incoming requests at a certain rate. Each server has a waiting queue where requests are queued while waiting for service. Since the service control nodes are central points of control, they are sensitive to overload. Therefore, admission control mechanisms are usually implemented in the nodes.

A good admission control mechanism improves the performance of the system during overload, whereas a bad mechanism may cause catastrophic results. Usually, server systems are analyzed with queueing theory. However, there are no queueing theoretic methods that can be used when developing and designing admission control mechanisms. Instead, control theory can be used. Control theory has since long been used to analyze different types of automatic control systems. Also, it contains a number of mathematical tools that may be used to analyze both the stability of a controlled system and to find good control schemes.

One well-known controller in automatic control is the PI-controller, which enables a stable control for most types of system (see, for example, [10]). Before designing the PI-controller, the system must be analyzed so that its dynamics during overload are known. Therefore, the system must be described with a control theoretic model. If the model is linear, it is easily analyzed with linear control theoretic methods. However, a queueing system is both non-linear and stochastic.

Very few papers have investigated admission control mechanisms for server systems with control theoretic methods. In [1] a web server was modelled as a static gain to find controller parameters for a PI-controller. A scheduling algorithm for an Apache web server was designed using system identification methods and linear control theory in [5]. However, the papers analyzing queueing systems with control theoretic usually describe the system with linear deterministic models. In [8] it is argued that deterministic models cannot be used when analyzing queueing systems. Until now, no papers have designed PI-controllers for server systems using non-linear control theory.

This paper investigates admission control mechanisms for a general service control node, which we model as a single server queue. We develop and analyze two control theoretic models, one non-linear and one linear. The non-linear model uses a non-linear fluid flow approximation first developed in [2]. The linear model includes the simplifications usually made when analyzing a queueing system with linear control theory. The main objective of the paper is to show the importance of using non-linear models when designing controllers for server systems using control theoretic methods. Further, the paper discusses some of the problems that may occur when linear control theory is used to analyze queueing systems.

## 2. Queueing model

In this paper, we assume that a service control node may be modelled as an M/G/1-system with an admission control mechanism, see Fig.1. New requests arrive according to a Poisson process with average rate $\lambda$ requests per second.

The objective of the controller is to keep the number of requests in the system, $x$, at a reference value, $x_{ref}$. Using $x$, the controller decides the rate, $u$, at which requests can be admitted to the system.

In the investigations we have used a PI-controller. The control law in continuous time is as follows:

$$u(t) = K \cdot e(t) + \frac{K}{T_i} \cdot \int_0^t e(v)dv$$

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

where *e(t)* is the error between the control variable and the reference value, that is $e(t) = x_{ref} - x(t)$. The gain $K$ and the integral time $T_i$ are the controller parameters that are set so that the controlled system behaves as desired. A large value of $K$ makes the controller faster, but weakens the stability. The integrating action eliminates stationary errors, but may also make the system less stable.

Continuous control is not possible in computer systems. Instead, time is divided into control intervals of length $h$ seconds. At the end of interval $k$, the controller calculates the desired admittance rate for interval $k+1$, denoted $u(k+1)$, by using the following control law:

$$u(k+1) = Ke(k) + \sum_{i=1}^{k} \frac{K}{T_{si}} e(i)$$

where $e(k) = x_{ref} - x(k)$. $x(k)$ is the number of requests in the system at the end of interval $k$. Since the controller is discrete, the controller parameter for the integration action, $T_{si}$, is given by $T_{si} = T_i / h$ where $T_i$ is the integral time in continuous-time.

The gate rejects those requests that cannot be admitted. The requests that are admitted proceed to the rest of the system. Since the admittance rate may never be larger than the arrival rate, $\lambda(t)$, the actual admittance rate, $\bar{u}(t) = \min[u(t), \lambda(t)]$. Admitted requests have a service time with mean value $1/\mu$ seconds.

In the investigations, the gate uses a leaky bucket algorithm to reject those requests that cannot be admitted. An arriving request is only admitted if there is an available ticket. New tickets are generated at a rate of $u(k)$ tickets per second during interval $k$. There can be maximum $H$ available tickets at a certain time.

The objective of this paper is *not* to show that service control nodes can be modelled as M/G/1-systems. Instead, our aim is to show the benefits of using non-linear control theory when designing admission control mechanisms for server systems. Therefore, the dynamics of an M/G/1-system are enough for this purpose.

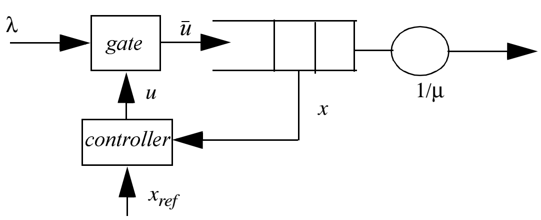


**Figure 1.** Investigated system.

## 3. Flow models for single server queues

When analyzing a queuing system with control theoretic methods, it is necessary to have a mathematical model that mimics the behavior of the system. One common approach is to use a flow approximation in which the arrival process is seen as a flow that enters the queueing system. Denote the average number of jobs in the system at time $t$ with $x(t)$. By writing the flow conservation equation, a queue may be described with the following differential equation:

$$\frac{dx}{dt} = \lambda(t) - \mu\rho(t) \tag{1}$$

where $\lambda(t)$ is the average arrival rate, $\mu$ is the average service rate, and $\rho(t)$ is the average utilization of the server at time $t$.

A linear flow model using the equations above has been analyzed in numerous articles, especially concerning ATM flow control (see, for example, [6]). In this model, $\rho(t)=0$ when $x(t)=0$, and otherwise $\rho(t)=1$.

A non-linear flow model was first developed by [2] and was further investigated in [9]. The non-linear flow model captures some of the stochastic behavior of a queueing system. It is shown that the model accurately mimics the behavior of a queueing system during non-stationary traffic conditions.

In this model, $\rho(t)$ is approximated by a non-linear function $G(x(t))$. $G(x(t))$ is found by assuming that at steady state, that is when $\dot{x} = 0$, the following relationship can be determined:

$$\bar{x} = F(\bar{\rho}) \tag{2}$$

where $\bar{x}$ is the average number of requests in the system and $\bar{\rho}$ is the server utilization at steady state. By assuming that $G(x(t)) = F^{-1}(x(t))$, (1) becomes solvable.

In [9] explicit expressions for $G(x(t))$ are developed for a number of well-known queueing systems. For an $M/G/1$-system, the expression becomes

$$G(x(t)) = \frac{x(t) + 1 - \sqrt{x^2(t) + 2C^2 x(t) + 1}}{1 - C^2} \tag{3}$$

where $C^2$ is the squared coefficient of variance of the service time distribution.

The main advantage of the non-linear model is that it captures some of the dynamics of a stochastic system. The approximated system is correct in terms of average number of customers and server utilization during steady state. The disadvantage is that the non-linear model is more difficult to analyze than the linear model.

## 4. Control theoretic models

In a control theoretic model, the system is described in terms of transfer functions or differential (or difference) equations. In this paper we develop and analyze two control theoretic models of an M/G/1-system with admission control. The models are based on the mathematical approx-

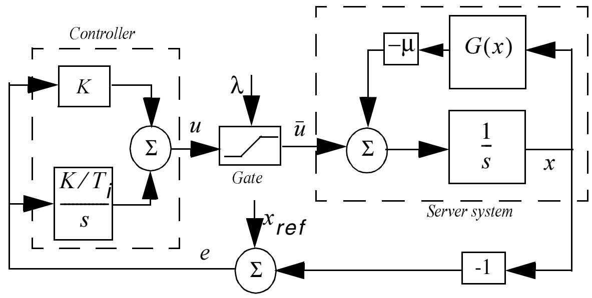


**Figure 2.** Non-linear control theoretic model.

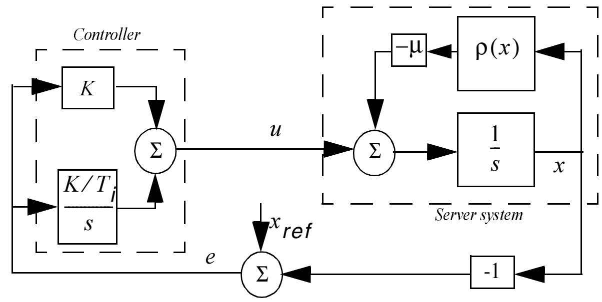


**Figure 3.** Linear control theoretic model.

imations described in section 3. The first model is non-linear and the second model is linear.

### 4.1. Non-linear model

A non-linear control theoretic model is shown in Fig. 2. The objective of the controller is to minimize the error between the number of requests in the system, $x(t)$, and a reference value $x_{ref}$. The Laplace transform of the control law for the PI-controller is given by

$$C(s) = K\left(1 + \frac{1}{T_i s}\right) \tag{4}$$

The gate saturates the control signal, $u(t)$, between zero and $\lambda(t)$. This means that we introduce two non-linearities when calculation the actual admittance rate. The rate may not be negative and it cannot be higher than the actual arrival rate. This is of course the case also in real server systems.

The server system is modelled as an integrator, representing the queue, and a non-linear feedback. The non-linear feedback in the server system is given by (3).

There are two main advantages with this model. First, several papers have shown that if the non-linear feedback in the server system is chosen appropriately, the model mimics the behavior of the corresponding queueing system with high accuracy. Also, this model is rather simple to analyze with non-linear control theoretic methods. The non-linear feedback is static and thereby easy to analyze.

One disadvantage with the model is that it may be difficult to find an accurate expression for the non-linear feedback. Further, the admission control mechanism changes the original arrival process during overload. This means that the optimal feedback term may depend on the current load in the system.

### 4.2. Linear model

Previous research analyzing queueing systems with control theoretic methods have usually developed linear models of the systems. Therefore, we here present a similar linear model of a service control node. In the investigations we compare this model with the non-linear model.

To describe a queueing system with admission control as a linear system three assumptions must be made. First, the system is assumed to be deterministic. This means that the the server system may be described as a simple integrator. Second, the arrival process is assumed to be unlimited and negative control signals are allowed. This means that the non-linearity in the gate is deleted. With an unlimited arrival process, the control signal cannot saturate due to the current arrival rate. Negative control signals means that the system may produce "negative" customers. When a negative customer enters the queue, the queue length is *decreased* with one.

The linear model is shown in Fig. 3. The feedback term, $\rho(x)$, equals *one* when $x$>0, and *zero* otherwise. Even if this term is non-linear, it is so simple that we have decided to keep it. Without this term, the queue length is allowed to be negative. Previous papers concerning for example ATM flow control have removed this non-linearity by assuming that the queue never is empty (see, for example, [6]).

This model is very simple to analyze analytically with linear control theoretic methods. However, one disadvantage with this model is that it assumes a deterministic behavior of the system, which means that the stochastic nature of a queueing system is ignored. This means, for example, that the average queue length becomes zero when $\lambda(t)<\mu$ for a longer time period. Another disadvantage is that the non-linearities introduced by the gate in a real system are ignored. The numerical investigations will show that this fact may cause problems when designing controller parameters for a server system.

## 5. Design of controller parameters

Before a PI-controller can be implemented in a server system, the controller parameters $K$ and $T_i$ must be determined. In this section we show how the parameters can be chosen by analyzing the non-linear model. The server system can be described with the differential equation

$$\frac{dx}{dt} = \bar{u} - \mu G(x(t))$$

Linearizing the system around $x_{ref}$ and $\overset{\circ}{\bar{u}} = \mu G(x(t))$, we get the linearized system

$$\frac{d\Delta x}{dt} = -\gamma \Delta x + \Delta \bar{u}$$

where $\gamma$ is given by

$$\gamma = \frac{\mu}{1-C^2}\left(1 - \frac{x_{ref} + C^2}{\sqrt{x_{ref}^2 + 2C^2 x_{ref} + 1}}\right)$$

Let the system be controlled with a PI controller with

$$\Delta\bar{u}(t) = K\left(e(t) + \frac{1}{T_i}\int e(\tau)d\tau\right)$$

where $e(t)=x_{ref}(t)-x(t)$. The closed loop system is in Laplace transform therefore given by

$$\Delta X(s) = \frac{K(s + 1/T_i)}{s^2 + (\gamma + K)s + K/T_i}\Delta X_{ref}(s) \tag{5}$$

Assume that the desired characteristic equation is

$$s^2 + a_1 s + a_2 = 0 \tag{6}$$

The values of the control parameters that gives this are

$$K = a_1 - \gamma \qquad T_i = \frac{a_1 - \gamma}{a_2} \tag{7}$$

Depending of the desired speed of the response of the closed loop system we can determine the coefficients in (6) and from (7) obtain the controller parameters.

A few comments could be made based on linear design methods: Poles located on the negative real axis will give non-oscillatory step-responses while complex conjugate poles will cause oscillatory responses. The farther into the left half-plane the poles are located, the faster the response will be in general. However, when we have saturations in the gate, we do not want to make the system too fast as an excessive desired control signal before the saturation will cause 'integrator windup' and worsen the performance (see [10]). The numerical investigations contains a discussion about anti-windup mechanisms.

Fig. 4 shows the closed loop poles for the linearized system. Two different designs, *D1* and *D2*, have their pole configurations depicted by 'x' and '*' respectively. According to the discussion above, *D1* is a "good" design and *D2* is a "bad" design.

## 6. Numerical investigations

In the numerical investigations we compare the control theoretic models with simulations of the corresponding queueing system. Through all investigations, $x_{ref}$ was set to

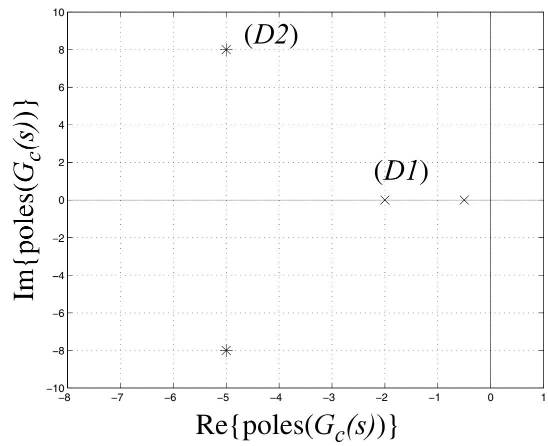


**Figure 4.** Pole locations for two different designs

10. The numerical results will be shown in the form of step responses.

The simulations were performed using a discrete-event simulation program implemented in C. In the simulations, the sample interval, *h*, was set to 0.5 seconds. The service times were hyper exponential distributed with parameters $\mu_1$=2 $s^{-1}$, $\mu_2$=60 $s^{-1}$, and $\alpha_1$=0.38, which meant that the average service time was 0.2 seconds and $C^2$=3.7. Once each second, the number of jobs in the system was measured, and it is these values that are presented in the graphs.

The results from the control theoretic models were produced using the Matlab Simulink package. In the models, the average service rate, $\mu$, was set to 5 and $C^2$=3.7.

### 6.1. Non-linear model

One of the objectives when using control theoretic analysis is to determine appropriate controller parameters for a given system. Our investigations show that the non-linear model can be used when designing the PI-controller. Due to lack of space, we will in this paper only show two examples of how the controller parameters can be designed, design *D1* and *D2* from the previous section. The average arrival rate, $\lambda$, is 20. This means that the system is heavily loaded, since the offered load is 4 Erlangs.

Fig. 5 shows the behavior of the system when the controller parameters are chosen well. We have used design *D1*, which means that *K*=2.4 and $T_i$=2.4. The step response for the simulations has a small overshoot in the beginning. However, both models have a short settling time and the variation in the queue length is relatively small. Both the average of 1000 realizations and the result of one realization are shown. As can be seen, the queue length varies during a single realization due to the statistical fluctuations in the system.

Fig. 6 shows the behavior of the system when the controller parameters are chosen badly. We have used design

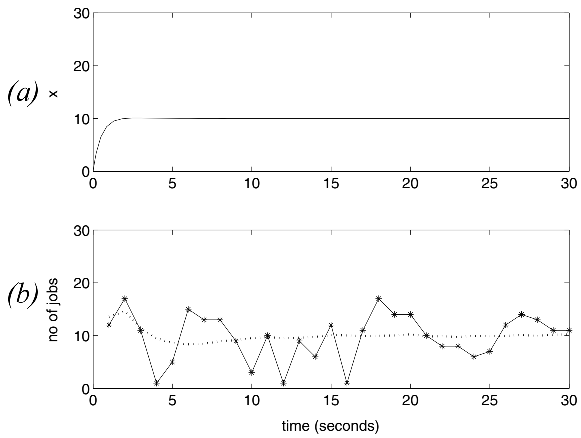


**Figure 5.** Good controller parameter design
(a) non-linear control theoretic model;
(b) simulations: -- average, -*- one realization

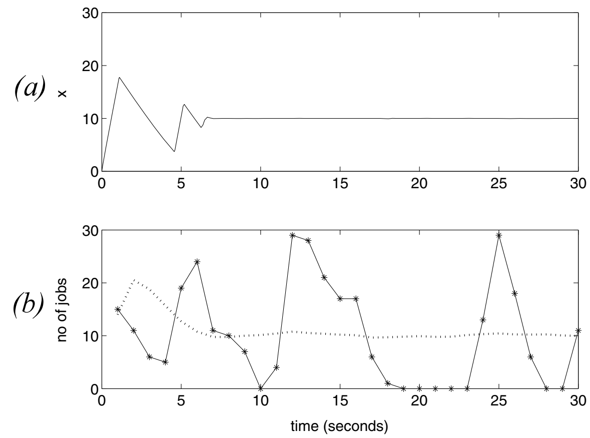


**Figure 6.** Bad design of controller parameters:
(a) non-linear control theoretic model;
(b) simulations: -- average, -*- one realization

*D2,* which means that $K$=9.9 and $T_i$=0.11.The step response for the non-linear control theoretic model oscillates for about 7 seconds before settling at the reference value. Even if the average step response for the simulations seems to be very smooth, the result for one realization shows that the system is oscillating.

### 6.2. Linear model

The linear model is much easier to analyze than the non-linear model. If we assume that the system is in the linear region, that is when the queue length is above zero, the closed loop transfer function from $x_{ref}$ to $x$ is given by

$$\frac{K(s + 1/T_i)}{s^2 + Ks + K/T_i}$$

which means that for a given characteristic polynomial, $s^2 + a_1 s + a_2 = 0$, the controller parameters are given by

$$K = a_1 \qquad T_i = \frac{K}{a_2} \tag{8}$$

This design looks very similar to the non-linear design, however, the two models behave very differently in some cases. In the linear model there are no saturations that can cause control problems. This means that the controller parameters may be chosen so that the linear step response becomes very fast.

One example of this is shown in Fig. 7. In this example λ=20. The poles of the controlled system are placed in -10 and -8, which means that $K$=18 and $T_i$=0.2. As can be seen in diagram (a), the linear model has a very good behavior, with a short settling time and only a small overshoot. The non-linear model, however, has some oscillations before settling at the reference value. Diagram (b) shows that the simulated system behaves very badly.

The oscillations in the non-linear model and the simulations are due to saturations in the gate. This saturation problem is inherent in the system when the arrival process is limited (i.e. greedy sources are not present). It is therefore of great importance to consider this non-linearity in the analysis as else the evaluation will be very misleading. Had the linear model been used in the parameter design of a real server system, the oscillations shown in Fig. 7 would not have been detected before implementation.

### 6.3. Anti-windup mechanisms

When a controlled system contains saturations, as the described system in this paper, the integrator action in the PI-controller may suffer from so called integrator windup. Usually, this problem is solved by implementing a so called anti-windup mechanism in the controller. The anti-windup mechanism calculates an error signal, $e_s$, given by

$$e \ = \bar{u}(t) - u(t) \tag{9}$$

where *u(t)* is the desired control signal and $\bar{u}(t)$ is the actual control signal to the system. If the actual control signal differs from the desired control signal, the anti-windup action tries to make the error signal zero. Thereby the integration action in the controller is reset. The control law for a PI-controller with anti-windup is given by

$$u(t) = Ke(t) + \int_0^t \left(\frac{K}{T_i}e(v) + \frac{1}{T_t}e\ (v)\right)dv$$

where $T$ is the anti-windup parameter. A typical choice is $T_t$=$T_i$.

Fig. 8 shows an example of the benefits using an anti-windup mechanism. In this example λ=10, $K$=18 and $T_i$=0.2. The step response for the non-linear model is shown both with and without anti-windup. In the case with anti-windup, $T_t$=0.2. As can be seen, the system with anti-win-

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

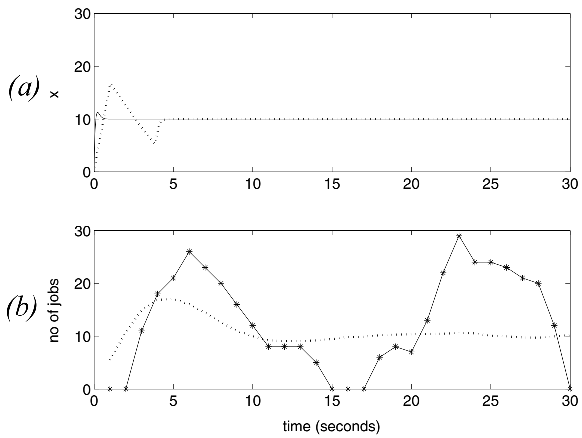


**Figure 7.** Design of controller parameters using linear model.
(a) solid line: linear model, dotted line: non-linear model.
(b) simulations: -- average, -*- one realization

dup has an almost perfect step response without the oscillations occurring in the system with no anti-wind up.

## 7. Conclusions

Control theory contains several mathematical tools useful when designing admission control mechanisms for server systems. However, before a server system may be analyzed with control theoretic methods, it must be described in terms of transfer functions or differential equations.

Previous papers have usually developed linear models when analyzing queueing systems with control theoretic methods. These models assume deterministic systems with unlimited arrival processes.

The main objective of this paper has been to show how non-linear control theory can be used when designing admission control mechanisms for server systems. In the numerical investigations it is shown that linear models are insufficient for these systems. A server system usually has a stochastic arrival process and non-linear dynamics. If a linear model is used when determining controller parameters, the real system may not behave as desired.

Therefore, we develop and analyze a non-linear control theoretic model of a server system. We show how a PI-controller may be designed using the non-linear model. The non-linear model is then compared with simulations of the corresponding server system.

## 8. Acknowledgements

The work in this paper has partially been supported by the Swedish Research Council through the Multi Project Grant 621-2001-3020 and contract 621-2001-3053.

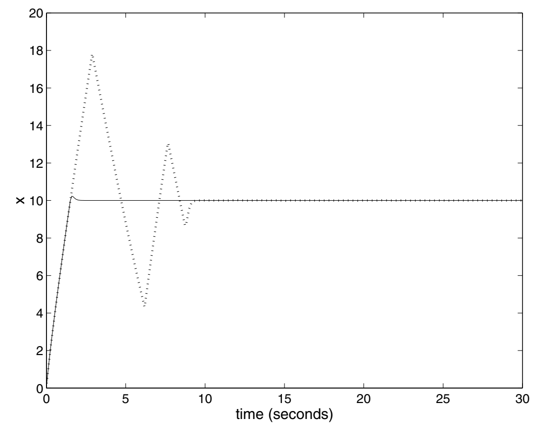


**Figure 8.** Non-linear model
dotted line: without anti-windup,
solid line: with anti-windup

## 9. References

[1] T.F. Abdelzaher, K.G. Shin and N. Bhatti, "Performance guarantees for web server end-systems: a control theoretic approach", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 13, No. 1, Jan 2002, pp 80-96.

[2] C.E. Agnew, "Dynamic modeling and control of congestion-prone systems", *Operations Research*, Vol. 24, No. 3, May-June 1976, pp 400-419.

[3] G.F. Franklin, J.D. Powell, A. Emami-Naeini and D.J. Powell, *Feedback control of dynamic systems*, Prentice Hall, 4th ed., 2002.

[4] L. Kleinrock, *Queueing Systems Vol.1*, John Wiley & Sons, 1975.

[5] C. Lu, T.F. Abdelzaher, J.A. Stankovic and S.H. Son, "A feedback control approach for guaranteeing relative delays in web servers", Proc. of the 7th IEEE Real-Time Technology and Applications Symposium, 2001, pp 51-62.

[6] S. Mascolo, "Smith's principle for congestion control in high-speed data networks", *IEEE Transactions on Automatic Control*, Vol. 45, No.2, Feb 2000, pp 358-364.

[7] J.E. Slotine and W. Li, *Applied nonlinear control*, Prentice Hall, 1991.

[8] S. Stidham Jr., "Optimal control of admission to a queueing system", *IEEE Transactions on Automatic Control*, Vol.30, No.8, Aug 1985, pp 705-713.

[9] D. Tipper and M.K. Sundareshan, "Numerical models for modeling computer networks under nonstationary conditions", *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 9, Dec 1990, pp 1682-1695.

[10] K.J. Åström and B. Wittenmark, *Computer-controlled systems, theory and design*, Prentice Hall International Editions, 3rd Edition, 1997.