

# The Applicability of Adaptive Control Theory to QoS Design: Limitations and Solutions

Keqiang Wu<sup>1</sup>

<sup>1</sup>*Electrical and Computer Engineering  
University of Minnesota  
Minneapolis, MN 55455, USA  
kqw@ece.umn.edu, lilja@ece.umn.edu*

David J. Lilja<sup>1</sup>

<sup>2</sup>*Honeywell Aerospace Electronic Systems  
3660 Technology Drive  
Minneapolis, MN 55418, USA  
haowei.bai@honeywell.com*

Haowei Bai<sup>2</sup>

## Abstract

*Due to the increasing complexity, the behavior of large-scale distributed systems becomes difficult to predict. The ability of on-line identification and auto-tuning of adaptive control systems has made the adaptive control theoretical design an attractive approach for quality of service (QoS) guarantee. However, there is an inherent constraint in adaptive control systems, i.e. a conflict between asymptotically good control and asymptotically good parameter estimates. This paper addresses these limitations via sensitivity analysis. The simulation study demonstrates that the adaptive control theoretical design depends on the excitation signal, environment uncertainty, and a priori knowledge on the system. In addition, this paper proposes an adaptive dual control framework for mitigating these constraints in QoS design. By incorporating the existing uncertainty of the on-line prediction into the control strategy, the dual adaptive control framework optimizes the tradeoff between the control goal and the uncertainty.*

## 1. Introduction

The widespread deployment of the advanced computer technology in business and industries has demanded the high standard on quality of service (QoS). For example, many Internet applications, i.e. online trading, e-commerce, and real-time databases, etc., execute in an unpredictable general-purpose environment but require performance guarantees. Failure to meet performance specifications may result in customer complaints, financial loss, or liability violations. To design a computer system with certain performance guarantees becomes increasingly important.

The traditional approach to designing a computing system with QoS guarantees has been to quantify hardware capability, software execution requirements, resource demands, and workload characteristic, then apply an appropriate combination of pre-run-time

analysis, admission control, and resource allocation algorithms to ensure that the system is not overloaded and that the desired performance is achieved. This approach is inherently feed-forward and open loop. During the past years, there have been several designs based on the conventional feedback control-theoretic theory for QoS guarantee [1][9].

In order to meet the design requirements, both the feed-forward and feedback approaches rely on an accurate model for the system behavior. Compared with feed-forward open loop approaches, fixed parameter controllers based on the conventional feedback theory can be designed to be somehow insensitive against process variations, noise, or disturbance. However, such controllers must, by nature, be conservative in the sense that the bandwidth of the closed loop system has to be decreased to reduce the influence of the variation in the process. Substantial changes in the process behavior can significantly degrade control loop performance [2].

Due to the increasing scale and complexity of distributed computing systems, their behavior becomes more unpredictable. For example [6], enterprise-scale storage systems are large (with capacities often in the order of 100s of TBs), distributed, and increasing heterogeneous, with constantly evolving hardware and software. The model parameters oftentimes depend on the hardware and software configuration. It is impractical to repeat system profiling every time a system upgrade. Moreover, the system behavior is oftentimes dependent on workload. Their workloads are complex consisting of multiple overlapping I/O streams with unpredictable request patterns. The dynamic changing behavior of the distributed computing systems imposes hurdles on fixed parameters feedback and feed-forward designs for QoS guarantee.

In the area of automation and control, adaptive control is an important way to handle system uncertainties. For example, the dynamics of an airplane change significantly with speed, altitude, angle of attack, and so on. The traditional constant gain, linear feedback can work well in one operating condition.

However, difficulties can be encountered when operating conditions change. An adaptive control system integrates the controller design with on-line recursive parameter estimation (system identification). As a result, the system controller can automatically adjust its parameters in response to changes in process and disturbance dynamics.

The parameter-varying characteristic of large computing systems makes the self-tuning adaptive control theory an attractive technique. Recently, the adaptive control theoretic design has been proposed for providing the QoS guarantees in large distributed environments. A case study was conducted in designing an adaptive controller for managing cache resources in QoS-aware servers [8]. More recently, an adaptive controller based on an on-line recursive least-square estimator was designed to control access requests to a shared storage [6]. The adaptive control theoretic design was suggested as a general approach to control QoS for large-scale distributed systems.

Undoubtedly, the adaptive control theory is a powerful technique and has been widely used for controlling non-linear systems with unknown and changing behavior in avionics industry, automobile control, etc. Traditionally, adaptive controllers are based on the separation of system identification, i.e. parameter estimation, and controller design. The uncertainty of estimation is not taken into consideration for the controller design, and the parameter estimates are used for designing a controller as if they were the real values of the unknown parameters. As a result, for the controller to be effective, the system identification process must be accurate and timely.

Unfortunately, there is an inherent conflict between identification and control in adaptive control systems – a conflict between asymptotically good control and asymptotically good parameter estimates [7]. To obtain good system information for identification it is necessary to perturb the process. Normally, the information about the system increases with the level of perturbation. On the other hand, the specifications of the control system are such that the output normally should vary as little as possible. There has been increasing interest in applying adaptive control theory to design computing systems in both academic and industry research recently, but this constraint has not been sufficiently addressed.

This work focuses on the applicability, limitations, and improvements of adaptive control theoretical design on computer systems. Based on the adaptive control framework and the experimental results for QoS guarantees in a proxy cache [8], we conduct a

simulation study. This paper makes the following contributions.

1. Demonstrate that the adaptive control theoretical design is excitation-dependent. If a workload does not possess sufficient excitation levels, the model prediction for a computing system does not converge to actual values.
2. Demonstrate that there is no guarantee on the accuracy and convergence of on-line prediction in an uncertain environment.
3. Demonstrate that the adaptive control theoretic design is dependent on the initial “guess” on the parameters for an unknown system. If the “guesses” deviate from the real values significantly, the model prediction does not converge.
4. This paper also proposes an adaptive dual control framework that optimizes the tradeoff between the control goal and the system identification.

The rest of the paper is organized as follows. Section 2 briefly reviews the conventional adaptive control framework [8]. Section 3 details the sensitivity analysis of on-line system identification. Section 4 briefly presents the dual adaptive control framework. Section 5 concludes.

## 2. Revisiting the Adaptive Control Framework for QoS Guarantees

### 2.1 The Framework Overview

Adaptive control systems are characterized by their ability to tune the controller parameters in real-time from the measurable information in the closed-loop system. Most of the adaptive control schemes are based on the separation of parameter estimation and controller design. This means that the identified parameters are used in the controller as if they were the real values of the unknown parameters.

A challenge on applying the adaptive control theory on computer system designs is to represent a computer system using “control language”. Previous research [6][8] typically assumed the system as a “black box”, whose current performance depends on a finite history of past measurement. Therefore, the system behavior can be described as a difference equation, whose parameters need to be identified. In order to model the web system behavior accurately, the parameter estimates must be accurate.

Figure 1 shows the structure of the adaptive control framework for QoS guarantee in a web cache system [8]. This model represents an application of adaptive control to provide proportional differentiation on relative average hit ratio of different content classes. In

generally, there are  $N \geq 2$  content classes in the system (for example, WML and regular HTML content). In a proportional differentiated caching service [9], the cache space is partitioned among classes, and assigning more storage space to a traffic class will increase its hit ratio and vice versa. Consequently, the quality spacing between classes is guaranteed by imposing constraints of the following form (equation 1) on successive pairs of classes, where  $H_i$  denotes the measured average hit ratio of  $class_i$  content, and  $S_i$  the allocated storage space for  $class_i$ .

$$H_i/H_j = S_i/S_j \quad (i, j=1, \dots, N) \dots (1)$$

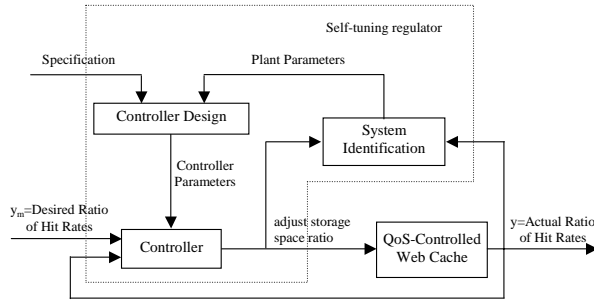


Figure 1. The adaptive QoS web cache control systems structure proposed in [8].

In the adaptive control web-caching system (Figure 1), an automatic model estimator (System Identification in Figure 1) periodically monitors actual system performance (relative hit ratio) and current resource allocation (storage space ratio). A mathematical model (i.e. estimated plant parameters) is derived based on the input-output relationship. This mathematical model is dynamically fed to a controller design block which in-turn design a controller based on the estimated model. It should be noted that the estimated parameters are used for designing a controller by assuming that they were the real parameters.

A system with two content classes ( $N=2$ ) was studied in [8]. An adaptive pole-placement controller was design to make the system output  $y(k)$  (which represents the measured relative hit ratio  $H_i/H_{j+1}$  at  $k$ th sampling time) track a reference trajectory  $y_m(k)$ . The task of the controller is to compute and provide the input  $u(k) = S_i/S_{j+1}$  to the proxy cache such that the control goal is achieved.

## 2.2 System Identification

The web cache system was modeled as a second order linear and time-invariant system (equation-2),

where the parameters  $p_0$ ,  $p_1$ ,  $r_0$ , and  $r_1$  are unknown and need to be estimated. These parameters represent the web cache system behavior. Once the system behavior changes, these parameters changes too. Therefore, it is important to estimate these parameters accurately. At every sampling time (e.g. 30 second), the input  $u(k)$  and output  $y(k)$  measurements are used for identifying these parameters, which is based on the on-line regressive least square estimation.

$$y(k) = -p_1 y(k-1) - p_2 y(k-2) + r_1 u(k-1) + r_2 u(k-2) \dots (2)$$

Two workloads, synthetic and empirical workloads, were used in their study. Their system identification results revealed that the change of traffic leads to the change of parameters, which implies that fixed parameter controller does not work well for all traffic patterns. Based on the on-line system identification, the system models under the synthetic and empirical workloads are described in equations (3) and (4), respectively.

$$y(k) = 1.13y(k-1) - 0.22y(k-2) + 0.0027u(k-1) + 0.0177u(k-2) \dots (3)$$

$$y(k) = 1.26y(k-1) - 0.34y(k-2) - 0.0082u(k-1) + 0.0517u(k-2) \dots (4)$$

## 2.3 Controller Design

Based on the on-line estimation of the web cache parameters, an adaptive controller was designed based on the pole-placement strategy. Pole-placement is a standard and widely-use rule for designing controllers. The pole location determines the system transient-response, such as speed, damping ratio, or bandwidth. The basic ideal was to determine a controller that gives desired closed-loop poles. Therefore, the system follows command signal in a specified manner. The adaptive pole-placement controller applied to the QoS proxy cache system can be represented in equation (5), where  $l$  is the controller order,  $a_{l-j}(k)$  and  $b_{l-j}(k)$ ,  $j=1, \dots, l$  are controller parameters that are automatically adjusted online. The detailed procedures for designing a pole-placement controller can be found in [2].

$$u(k) = \sum_{j=1}^l a_{l-j}(k) u(k-j) + \sum_{j=1}^l b_{l-j}(k) [y_m(k-j) - y(k-j)] \dots (5)$$

At every sampling time, the adaptive controller is fed with the output  $y$ , the reference  $y_m$ , and the plant input  $u$ . The controller computes and produces the new plant input  $u$  for the next sampling time. If the estimated parameters are accurate, the system output  $y$  should asymptotically track the reference  $y_m$ . However, the uncertainty of the parameter estimate is not incorporated into the controller design.

## 3. System Identification Sensitivity

To design an effective controller that can track variations in system behavior, an accurate system model is essential. Therefore, the on-line system identification becomes crucial. Two criteria determine the effectiveness of the on-line system identification, i.e. whether the parameter estimates converge the actual values, and how fast the estimates converge. There are several practical issues that have a significant impact on these two criteria. While the adaptive control theory has been applied on QoS design in previous studies [6][8], these issues have not been addressed.

In this section, through a series of sensitivity experimental study on the on-line system identification technique, we demonstrate the limitations of the adaptive control technique.

### 3.1 Experimental Setup

We develop a digital model (Figure 2) for the differentiated proxy caching service [8]. This model consists of system identification only. Under different workloads, the plant displays different behavior. The plant dynamics is based on the results reported in [8]. We also implement the same on-line system identification algorithm, i.e. Regressive Least Square Estimation [2]. For on-line estimation, an initial ‘guess’ on the system order and the system parameters  $[p_1, p_2, \dots, r_1, r_2, \dots]$  is required. If not indicate explicitly in the following experiments, the system behavior is assumed to be a second order system and the estimates for the system parameters  $[p_1, p_2, r_1, r_2]$  are initialized to some arbitrary values within  $[-1, +1]$ . These set-ups are similar to [8].

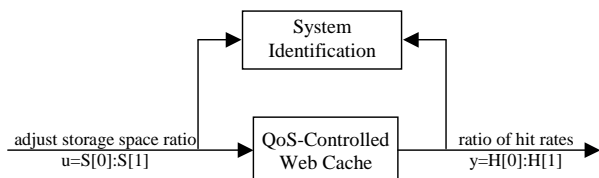


Figure 2. An on-line least square estimator for identifying a differentiated proxy caching system.

In addition to the two workloads (synthetic trace and empirical trace) used in [8], we introduce a workload with uncertainty, stochastic workload. It is assumed that, under the stochastic workload, the web cache displays uncertain behavior as described by equation (6), where the parameter drift is represented by  $a(k)$ , a white noise drift vector with zero mean and a small variance. This workload represents the system dynamics transition between synthetic and empirical

workloads but with some uncertainty. The variance magnitude determines the significance of uncertainty of the web cache behavior.

$$y(k)=[1.195+\varepsilon_1(k-1)]y(k-1)-[0.28+\varepsilon_2(k-1)]y(k-2) \\ -[0.0055+\varepsilon_3(k-1)]u(k-1)+[0.0347+\varepsilon_4(k-1)]u(k-2) \dots(6)$$

We compare the estimator performance using four types of excitation signals. Table I summarizes the three workloads and four excitation signals. An excitation is defined as a signal applied at input to the web cache system, i.e. storage space ratio in this study.

TABLE I. SUMMARY OF THE THREE WORKLOADS AND THE FOUR EXCITATION SIGNALS.

Workload	Synthetic trace
	Empirical trace
	Stochastic trace with variance as 0.0001 and 0.01
Excitation $u$ (storage space ratio signal)	Ex-1: $4\cos(2t)$
	Ex-2: $\cos(t)+\sin(2t)+\cos(3t)+\sin(5t)$
	Ex-3: white noise with variance 0.1
	Ex-4: white noise with variance 1.0

### 3.2 Sensitivity Analysis

In this section, we discuss the impact of excitation signal, workload uncertainty, and *a priori* knowledge on the system behavior. Since the web cache system behavior is represented by these parameters, whether the parameter estimates can converge to the actual values in a timely manner becomes critical.

**3.2.1 Impact of Excitation.** Whether the on-line estimator can predict the web cache behavior is the key to the adaptive controller performance. Therefore, the on-line parameter estimation must be accurate and on time. However, whether the on-line estimation can converge relies on the excitation signal, i.e. the storage space ratio  $u$ , to the web cache.

We conduct two sets of experiments with the workloads as synthetic trace and empirical trace, respectively. For each set of experiments, we use four different excitations as shown on Table I. Figure 3 (a to d) shows the estimated parameters for the synthetic trace workload under the four excitations. For Ex-1 and Ex-2, the estimations for all parameters significantly deviate from the actual values. For example, the estimate for  $r_2$  asymptotically approaches to 50 and 18 for Ex-1 and Ex-2 respectively, which are more than 1000 times for than the actual value, 0.0177 (Figure 3-d). For Ex-3, the estimated parameters changes slowly, and after 5000 seconds they all converge to the plant parameters as reported in [8].

When the variance of the white noise (represent the signal power) is increased, the estimated parameters in Exp-4 converge at a faster rate, i.e. all estimates converge after 3000 seconds. Similar patterns are observed for the estimated parameters under the empirical trace workload.

Obviously, it is impossible to design an effective controller based on the estimated parameters under the excitation 1 and 2. The comparison indicates that (i) whether the estimated parameters converge is excitation-dependent and (ii) the frequency contents of an excitation are more important than the magnitude of the excitation.

While previous studies [6][8] show success on on-line system identifications, it is important to realize that the excitation signal must be persistently exciting or sufficiently rich [2][3]. Unfortunately, there is an inherent conflict between identification and control in adaptive control — *a conflict between asymptotically good control and asymptotically good parameter estimates* [7]. To obtain good system information for identification it is necessary to perturb the process. Normally, the information about the system will increase with the level of perturbation. On the other hand the specifications of the closed loop system are such that the output normally should vary as little as possible. Since the excitation signal to the web cache is

generated by the controller which takes the feedback as input, good control may lead to a lack of identifiability due to a poor excitation. As a result, there is no guarantee that the web system will be properly excited.

**3.2.2 Impact of Workload Uncertainty.** The workload trace behavior in reality consists of multiple overlapping I/O streams and is dynamically changing with unpredictable request pattern [6][8]. We model this uncertainty by using the stochastic workload (Table 1). The variance represents the significance of the uncertainty. Since under the excitation of Ex-4, the on-line parameter estimator displays the best performance. In this section, we examine the on-line estimator performance under the stochastic workload with Ex-4 as the excitation.

Figure 4 (a to d) shows that, when the uncertainty is small ( $var=0.0001$ ), after 2000 seconds, the estimated parameters converge to their actual values. However, when the uncertainty increases ( $var=0.01$ ), even after 6000 seconds, the estimated parameters do not converge to their actual values. For example, the estimated parameter  $r_1$  deviates from the actual value by 30% even after 4000 seconds.

**3.2.3. Impact of Initial Guess on System Behavior.** Similar to [8], we have initialized the parameters  $[p_1,$

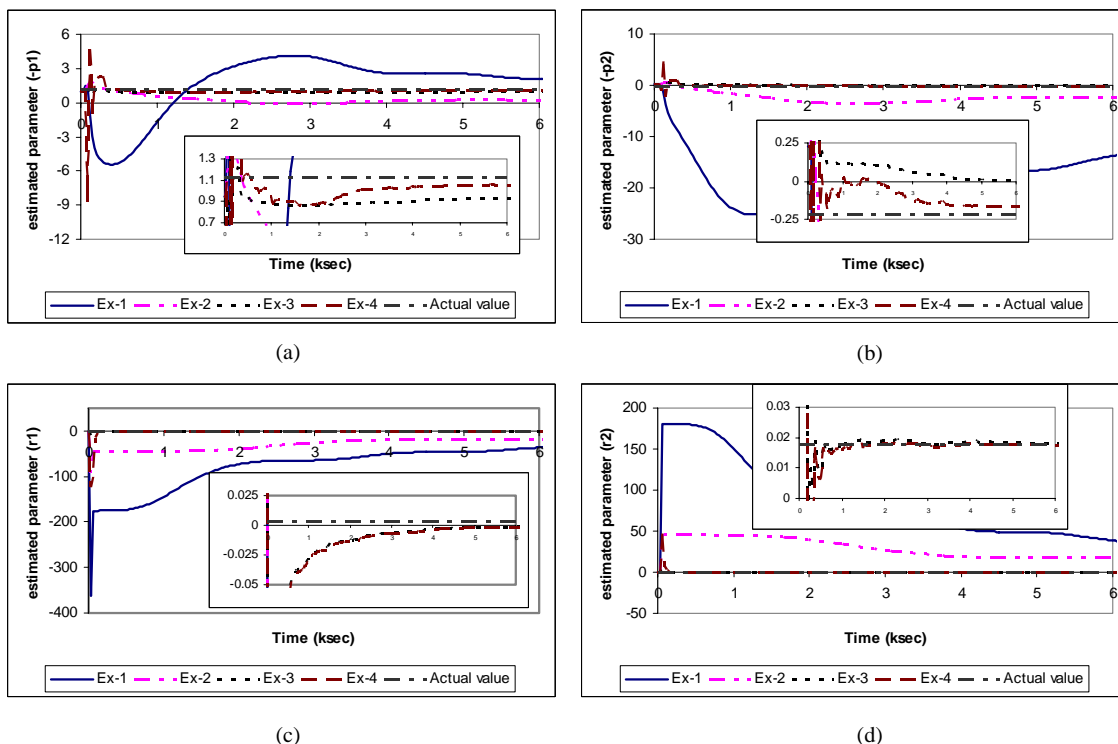


Figure 3. The dynamics of estimated parameters with the synthetic trace.

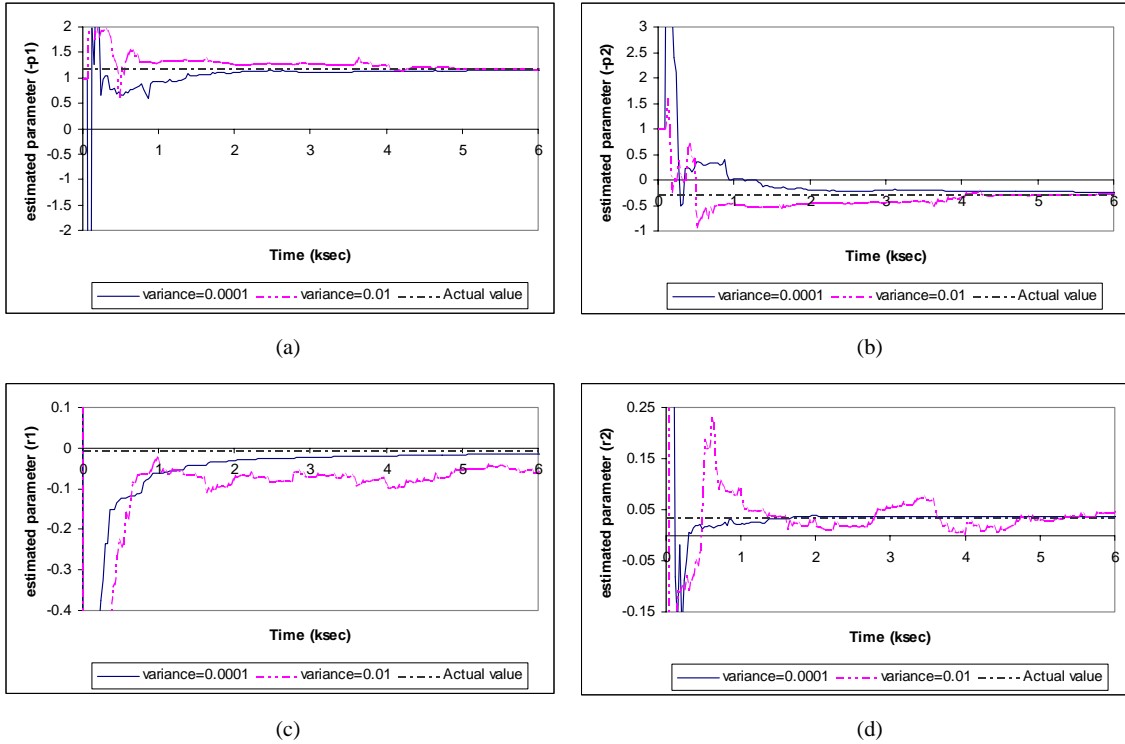


Figure 4. The dynamics of estimated parameters with the stochastic trace.

$p_2, r_1, r_2]$  to be some random values within  $[-1, +1]$ , which do not deviate too much from the actual parameters. In this section, we relax these constraints and examine the impact that if *a priori* knowledge is not available. We use Ex-4 as the excitation input.

Figure 5 (a to d) shows that, under the synthetic trace with the initial parameters guess as some random values within  $[-10, +10]$ , the estimated parameters (such as  $p_1, p_2$ , and  $r_2$ ) deviate from the actual values even after 6000 seconds. Under a large-scale distributed environment, it is more practical to assume that we do not have close *a priori* knowledge of the system behavior. This constraint reduces the prediction accuracy and increases the time for prediction to converge.

**3.2.4 Summary.** The original motivation of applying adaptive control theoretical design to computing systems is to automate the performance tuning process. One of the critical factors is whether the on-line parameter estimation can be accurate and on time. The sensitivity analysis in this section indicates that sufficient excitation and good *a priori* knowledge of the system behavior and workload characteristic are important for achieving accurate and on-time estimate. However, there is a dilemma between asymptotically good control and asymptotically good parameter

estimate. In addition, for a large-scale distributed computing system, the workload, as well as the system behavior, is dynamically changing. Good *a priori* knowledge might not be available.

As a result, to design a control strategy for a computing system that operates under uncertainty conditions should incorporate the existing uncertainty. The control signal should have the following properties: (i) it cautiously follows the control goal, which means, in the case of uncertainty parameters of the system, the control signal should be smaller (cautious) than the control signal in the system with known parameters; and (ii) after adaptation, it excites the plant to improve the estimation.

#### 4. An Adaptive Dual Control Framework for QoS Guarantees

In this section, we introduce a technique, adaptive dual control theory [4][5], for mitigating the inherent constraint of adaptive control theory, the dilemma between asymptotically good control and asymptotically good parameter estimates. The basic idea of dual control theory is to incorporate the existing uncertainty in the control strategy with the control signal. First the controller must control the process as well as possible. Second, the controller must

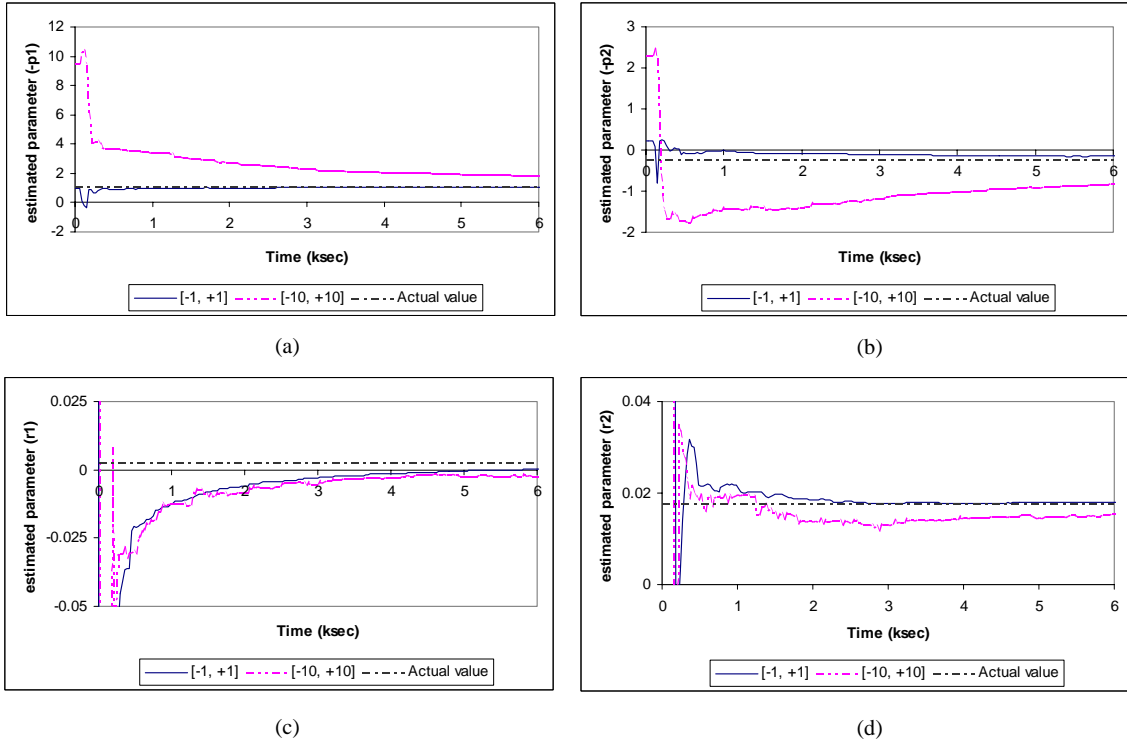


Figure 5. The dynamics of estimated parameters with the initial parameter guess within different ranges.

inject a probing signal or perturbation to get more information about the process. By gaining more process information, better control can be achieved in future time. The implementation of this technique is in process.

Figure 6 shows the structure of the adaptive dual control framework for the QoS web cache system. The main differences between an adaptive dual control system and a conventional adaptive control system (Figure 1) are the transmission of the accuracy of the parameter estimates from the estimation to the control design algorithm, and the combination of caution control and excitation. The utilization of the accuracy of the estimation for the controller design allows generating the optimal excitation and cautious control signal for an adaptive dual controller.

While the web cache system was modeled as a *linear time-invariant* system [8], we model the system as a discrete system with *time-varying parameters* (equation 7). Considering an enterprise-scale web cache system is larger, distributed, and increasing heterogeneous, with constantly evolving hardware and software, and its behavior also depends on workloads that consists of multiple overlapping I/O streams with unpredictable request pattern, it is more reasonable to model the system with time-varying parameters.

$$y(k+1) = -a_1(k)y(k) + \dots - a_n(k)y(k-n+1) + b_1(k)u(k) + \dots + b_m(k)u(k-m+1) \quad \dots(7)$$

where  $y(k)$  is the actual ratio of hit rate,  $u(k)$  the control signal for adjusting storage space ratio,  $k$  is the discrete time index;  $a_i(k)$  and  $b_j(k)$  for  $i=1, \dots, n$ , and  $j=1, \dots, m$  are the unknown time-varying system parameters. Equation (7) can be put in vector form as

$$y(k+1) = \mathbf{p}^T(k)\mathbf{m}(k) \quad \dots(8)$$

Where

$$\mathbf{p}(k) = [-a_1(k) \dots -a_n(k) | b_1(k) \dots b_m(k)]^T \quad \dots(9)$$

and

$$\mathbf{m}(k) = [y(k) \dots y(k-n+1) | u(k) \dots u(k-m+1)]^T \quad \dots(10)$$

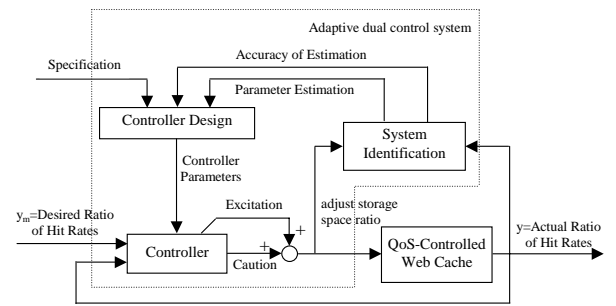


Figure 6. The adaptive dual control framework for QoS web cache systems.

We model the uncertainty by using an additional stochastic parameter drift (equation 11).

$$\mathbf{p}(k+1)=\mathbf{p}(k)+\boldsymbol{\varepsilon}(k) \dots(11)$$

where the noise vector  $\boldsymbol{\varepsilon}(k)$  is a white noise drift vector with zero mean and a small variance. The parameter vector  $\mathbf{p}(k)$  is estimated using the standard technique for on-line system identification.

There are two major issues in designing an adaptive dual controller, i.e. (1) selecting an appropriate performance index for control optimization; (2) describing the uncertainty of the controller parameters in the adaptive pole-placement control system as well as defining a measure for the uncertainty. For easy understanding, we avoid the detail process of mathematical derivation. Readers can refer [4][5] for the detail.

We define the nominal output as the system response to the controller that provides the desired system dynamics when no uncertainty exists. It is clear that the control performance would be improved if, in case of disturbances and parameter uncertainties, the controller tried to bring the system output as close as possible to the nominal output after complete noise compensation. The following two cost functions to be minimized are introduced in order to derive the control law, where  $\mathfrak{S}_k$  is the set of input and output values available at time  $k$ .

$$J_k^a = -E\{\beta^2[y_n(k+1)-y(k+1)]^2|\mathfrak{S}_k\} \dots(12)$$

$$J_k^a = -E\{[y(k+1)+\sum_{i=1}^m c_i y(k-i+1)-\hat{\mathbf{p}}^T \mathbf{m}(k)]^2|\mathfrak{S}_k\} \dots(13)$$

The first cost function, equation (12), is used for control purposes to minimize the deviation of the system output  $y(k+1)$ , from the unknown nominal output  $y_n(k+1)$ , which would be obtained by the adjusted unknown regulator. The coefficient  $\beta^2$  is introduced for the simplification of further algebraic manipulations. The second cost function, equation (13), is used for the acceleration of the parameter estimation process by increasing the predictive error value, where  $\hat{\mathbf{p}}$  is an estimate of  $\mathbf{p}$ , and  $c_i$  is determined the desired pole values. These two criteria correspond to the two goals of adaptive dual control: *to control the system output and to accelerate the estimation for future control improvement*. The adaptive dual controller is designed by solving this optimization problem (minimization of equations (12) and (13)).

## 5. Conclusions and Future Work

In this paper, we have addressed the limitations of applying the adaptive control theory to design computing systems with QoS guarantee via sensitivity analysis. The effectiveness of the adaptive control technique on the computing system design relies on the excitation signal, *a priori* knowledge of the system behavior, and environment uncertainty. We propose an adaptive dual control framework for QoS guarantee for resolving these constraints. By incorporating the existing uncertainty of the on-line prediction into the control strategy, the dual adaptive control framework optimizes the tradeoff between the control goal and the uncertainty.

We are in the process of examining the performance of the adaptive dual control framework.

## 7. Acknowledgment

This project was supported in part by the Minnesota Supercomputing Institute and Honeywell Aerospace Electronic Systems.

## 8. References

- [1] T. Abdelzaher, K. Shin, and N. Bhatti. Performance guarantees for web server end-systems: A control-theoretical approach. IEEE Transactions on Parallel and Distributed Systems, June 2001.
- [2] K. J. Astrom and B. Wittenmark. Adaptive Control, Addison Wesley, 2nd edition, 1995.
- [3] K. J. Astrom. Theory and Application of Adaptive Control – a Survey. *Automatica*, vol. 19, pp.471-486, 1983.
- [4] Elliott. Direct adaptive pole placement with application to nonminimum phase systems. IEEE Trans. Autom. Control, 27, 720-722, 1982.
- [5] N. M. Filatov and H. Unbehauen. Adaptive Dual Control: Theory and Applications. Lecture Notes in Control and Information Sciences, Springer Verlag, 2004.
- [6] M. Karlsson, C. Karamanolis, and X. Zhu. Triage: Performance Isolation and Differentiation for Storage Systems. In International Workshop on Quality of Service (IWQoS), Montreal, Canada, June 2004.
- [7] P. R. Kumar and P. Varaiya. Stochastic Systems: Estimation, Identification and Adaptive Control, Prentice Hall, Englewood Cliffs, N. J., 1986.
- [8] Y. Lu, T. Abdelzaher, C. Lu, and G. Tao. An adaptive control framework for QoS guarantees and its application to differentiated caching services. In International Workshop on Quality of Service (IWQoS), pages 23--32, Miami Beach, FL, May 2002.
- [9] Y. Lu, A. Sexana, and T. Abdelzaher. Differentiated caching services; a control-theoretical approach. In Proceedings of the 2001 International Conference on Distributed Computing Systems, pages 615--622, 2001.