Improved Timing Control for Web Server Systems Using Internal State Information

Xue Liu, Rong Zheng, Jin Heo, Lui Sha
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, 61801
(01)217-333-3722
{xueliu, zheng4, jinheo, lrs}@cs.uiuc.edu

ABSTRACT
How to effectively allocate system resource to meet the Service Level Agreement (SLA) of Web servers is a challenging problem. In this paper, we propose an improved scheme for autonomous timing performance control in Web servers under highly dynamic traffic loads. We devise a novel delay regulation technique called Queue Length Model Based Feedback Control utilizing server internal state information to reduce response time variance in presence of bursty traffic. Both simulation and experimental studies using synthesized workloads and real-world Web traces demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors
C.4 Computer Systems Organization: Performance of Systems: Design Studies

General Terms
Design, Experimentation, Measurement

Keywords
Web Server, Queueing Model, Control Theory, SLA, Feedback

1. INTRODUCTION
Web server systems have become an integral part of our society. One major problem web hosting companies face is how to meet the Service Level Agreements (SLAs) with their clients without excessively over-provisioning resources. SLAs are usually expressed in the form of the average response time guarantee, above which is not acceptable to the clients. It is of great theoretical and practical interests to provide delay service guarantees to clients in Web server systems.

To control the average response time in Web servers, a Queueing Model Based Feedback Control architecture is proposed in [1]. However, we observe that the performance of Queueing Model Based Feedback Control deteriorates in presence of bursty Web traffic. In particular, clients may experience large variances in response time. In this paper we devise a new control approach, Queue Length Model Based Feedback Control to achieve better timing control performance. It utilizes server internal state information of queue length to better handle the transient behaviors caused by rapidly changing traffic loads. As a result, response time's variance is greatly reduced while its mean still meets delay reference.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

The rest of the paper is organized as follows. In Section 2, we identify the problem of previous Queueing Model Based Feedback Control. In Section 3, the new timing control approach i.e., Queue Length Model Based Feedback Control is proposed. In Section 4, we evaluate the performance of the proposed control approach through simulation and experimental studies using both synthetic and trace based workloads. Finally, we conclude the paper in Section 5.

2. PROBLEMS WITH QUEUEING MODEL BASED FEEDBACK CONTROL
Studies reveal that the Web traffic is bursty and exhibits self-similar properties [2]. It has been shown the burstiness of Web request traffic can be modeled using the Pareto On/Off distribution.

To evaluate the performance of different control approaches, we developed a Web server simulation package using the network simulator ns-2. In the simulation, there are three adjustable parameters to control the "burstiness" of the Pareto On/Off traffic. The *Burst_Time* corresponds to the mean length of On period of the traffic. The *Idle_Time* corresponds to the mean length of the Off period and *Interval* is the interarrival time during On periods.

Figure 1 shows the delay experienced by a single client connection with *Burst_Time=1*, *Idle_Time =10* and *Interval=0.1*

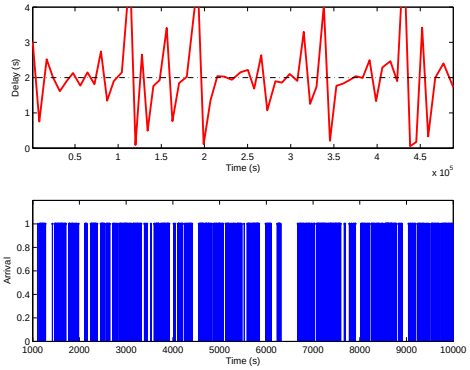

Figure 1: Performance of Queueing Model Based Feedback Control under a Single Pareto On/Off

under the original Queueing Model Based Feedback Control. The reference delay is set to $D^{ref}=2$. Each point in the upper figure is an average of response times experienced by *1000* requests. The corresponding mean and variance of the response time are *2.0041* and *1.2034* respectively. The bottom figure shows the request arrival versus the time line. We can see the traffic is indeed very busty and has interleaved On/Off periods.

From Figure 1, we observe that using the previous approach, the delay fluctuates a lot around the reference value although the long-time average mean delay is close to D^{ref}. This is because the online estimation of the request rate λ is a time average of the instantaneous request rate (on the order of *500* requests in our implementation). For bursty traffic, the instantaneous request rate is larger than the long-term average rate λ during On periods. However, the feed forward predictor's output rate is based on λ and thus is lower than the instantaneous request rate. Therefore, the request queue will build up and clients will experience longer response time. On the other hand, during periods with sporadic requests, the instantaneous request rate is smaller than λ, which leads to smaller response time. This observation motivates us to consider the use of server internal state information in the controller design to suppress large delay variations.

3. QUEUE LENGTH MODEL BASED FEEDBACK CONTROL

To reduce the response time variance under bursty traffic, we propose using server internal queue length measurements to predict service rate allocation μ_q. The procedure of the *Queue Length Model Based Predictor* is as follows:

Step 1: At each control invocation, we measure the current queue length $l_{current}$ and update the request rate estimate λ.

Step 2: Based on results from queueing theory, a targeted queue length $l_{targeted}$ is computed. For example, if *M/M/1* model is used to model the server's internal queue, we have $l_{targeted} = \lambda D^{ref}$.

Step 3: Let $\mu_q = (\frac{1}{D^{ref}} + \lambda) + K \times (l_{current} - l_{targeted})$ to be the new model output service rate. The first term $(\frac{1}{D^{ref}} + \lambda)$ is the same as the Queueing Model Predictor in the original approach. The second term $K \times (l_{current} - l_{targeted})$ represents the queue length feedback. K is a constant control gain, in practice, we can set $K = 1/D^{ref}$.

The proposed Queue Length Model Based Feedback Control regulator is shown in Figure 2.

4. EFFECT OF QUEUE LENGTH MODEL BASED FEEDBACK CONTROL

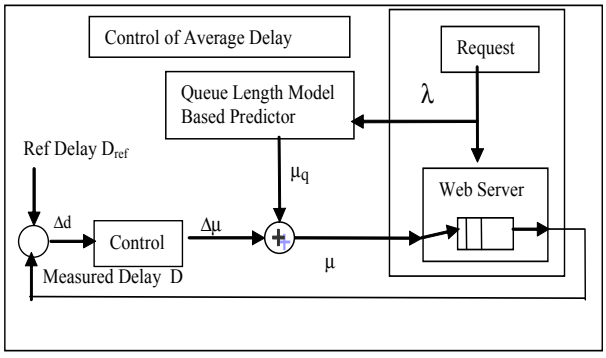

Figure 2: Queue Length Model Based Feedback Control

We first implemented the Queue Length Model Based Feedback Control in our simulation package. To verify the design of the proposed controller, we experiment with the same Pareto On/Off traffic (with parameters *Burst_Time = 1, Idle_Time = 10 and Interval = 0.1*) as in Section 2. Figure 3 demonstrates the controlled server performance using the new Queue Length Model Based Feedback Control. The variance of the client response time is *0.0051* as compared with *1.2034* using the previous Queueing Model Based Feedback Control approach. In addition, the average response time of both approaches are very close to $D^{ref} = 2$. We see using the new controller, even with this extremely bursty traffic, good delay regulation can be achieved.

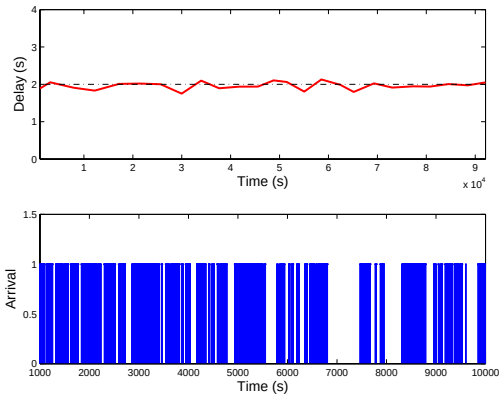

Figure 3: Performance of Queue Length Model Based Feedback Control under Pareto On/Off Source

We also implemented Queue Length Model Based Feedback Control on Apache web server 2.0.7 with Linux kernel 2.4.20. We tested the new approach under both synthetic and trace based workloads [3]. Due to space limit, we have to omit the results here; interested readers are referred to [3].

5. SUMMARY AND CONCLUSIONS

We proposed a new Web server timing control scheme called Queue Length Model Based Feedback Control. Compared with previous approaches, the new scheme can significantly reduce response time variance under a wide range of workload conditions including bursty traffic. This is achieved by utilizing the server internal queue length measurements. Extensive simulation study shows that the new scheme can provide smooth performance control and better track SLA specifications in Web server systems.

6. REFERENCES

[1] L. Sha, X. Liu, Y. Lu, T. Abdelzaher, "Queueing Model Based Network Server Performance Control", *IEEE Real-Time Systems Symposium*, Phoenix, Texas, Dec, 2002

[2] M. Crovella, A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Cause", *Proceedings of SIGMETRICS* 1996

[3] M. Arlitt and T. Jin, *1998 World Cup Web Site Access Logs*, Aug. 1998. Available at http://www.acm.org/sigcomm/ITA/

[4] X. Liu, R. Zheng, J. Heo and L. Sha, "Timing Performance Control in Web Server Systems Utilizing Internal State Information", extended version, Available at http://www-sal.cs.uiuc.edu/~xueliu/Timing.pdf