

**THURSDAY
FEBRUARY 5, 2004**

**Scaife Hall Auditorium
Room 125**

**4:00 PM
Refreshments—3:30 PM**



Lui Sha

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Lui Sha obtained his PhD from CMU in 1985. From August 1986 to July 1998, he was a Senior Member of Technical Staff at the Software Engineering Institute at CMU. Since the fall of 1998, he has been professor of Computer Science at UIUC.

He received the Outstanding Technical Contributions and Leadership award from the IEEE Technical Committee on Real Time Systems in Dec. 2001, and was elected to be a Fellow in 1998 "for technical leadership and research contributions, which enabled the transformation of real-time computing practice from an ad hoc process to an engineering process based on analytic methods." His research was reported in the Selected Accomplishments section of the National Research Council's report, A Broader Agenda for Computer Science and Engineering. He has been credited for making critical contributions to the success of many national high technology projects, including the GPS software upgrade, the Space Station, and the Mars Pathfinders.

Currently, he is active in the development of robust real time software and is a member of National Academy of Science's study group on software dependability and certification.

QUEUING MODEL BASED PERFORMANCE CONTROL

Network based server systems (e.g., Web servers) have now become an integral part of our society. Controlling the timing performance of each individual connection to a network server presents a challenging theoretical problem with important practical implications to web hosting companies, as well as operators of command and control systems.

In this talk, I will present a queuing model based feedback control approach to keep the timing performance of a network server close to the specification. From the perspective of QoS management, this is the dual problem of traditional approaches where what is explicitly controlled is specification on the resource usage, rather than the performance experienced by users.

Feedback control is an important technology for networked systems. For example, feedback control is embedded in the TCP protocol in the form of a sliding window mechanism. Since it was introduced in the 70's to solve the congestive failure problems that had brought down the network, we have not experienced system-wide congestive failures again even though the network has grown orders of magnitude. This is a testament of the effectiveness of feedback control in a highly dynamic, decentralized, and fast changing environment. However, going beyond congestion control to provide smooth performance control over a wide range of workload conditions is challenging. A network server's response to allocated resources is highly non-linear. In addition, the workload is stochastic and its parameters could change abruptly over a wide range of values.

It has been observed by many that fixed controllers tend to produce mediocre performance, except in the limited case of heavy workload that allows for reasonably accurate fluid approximation. However, a direct application of adaptive control or hybrid control does not yield significant performance improvements. The difficulty originates from the fact that the equilibrium states in a queuing system are as fickle as Web surfers' attention spans. In addition, the random fluctuations in arrivals, queue lengths, and response times make it time consuming in the estimation of the equilibrium states. Finally, the impact of the control itself can alter the equilibrium states. Together, they make it a challenging task to fine tune controllers as a function of the equilibrium states in a queuing system.

Queuing models are a "natural" to model the non-linear behaviors of a networked server over large ranges of parameter values. Thus, we can use online solutions from a queuing model as feed forward control to "lock" a queuing system into a desired equilibrium state, in spite of abrupt changes in workload. We can then design a controller aiming at suppressing the random fluctuations in response time within the given equilibrium state. This is a much easier task for a controller to handle. In addition, the controller corrects bias due to approximation errors in the queuing model. This method has been validated in Apache Web servers in the labs. It is intellectually satisfying to see that two separately developed powerful theories can work synergistically in a unified framework.

For more information:

<http://www.ece.cmu.edu/seminar/index.php>

Bruce Krogh, ECE Seminar Host
krogh@cmu.edu