Preface

In a computer system there is a wide spectrum of actions that are to be performed simultaneously in order to maximize throughput and minimize the response time of the system. The diversity of action and reaction which takes place in a large computer system presents serious problems to anyone trying to encompass that entire spectrum. No single theory of optimization – finding the best way to carry out an action to produce a specific result – embraces all the important aspects because each domain of application has unique characteristics that determine the most appropriate approach. There are many problems of long-standing interest which have eluded solution, partly because the problems have not been clearly defined and partly because the most significant parameters to be used in tractable models have not been identified. The eight papers in this issue introduce some important and novel concepts in the analysis and design of computer systems. New viewpoints have been developed for the analysis of significant aspects such as multiprocessing, multiprogramming, system communication, storage management, scheduling, and validation of simulation models. The concepts presented in this group of papers contain numerous warnings and restrictions, which serve to call attention to those aspects of the discussion that must be modified in order to represent in a more realistic manner the physical processes of existing systems. Success in the application of the described models depends on careful formulation by the user and an understanding of the assumptions and restrictions involved in their use.

Fernández and Lang propose a computation of lower bounds for multiprocessor schedules. The authors treat two basic problems: 1) A deadline for the execution time of a given set of tasks must be satisfied by a minimum number of processors; and 2) a fixed number of processors must be used to execute a set of tasks in a minimum time. Their assumptions are that the multiprocessing system is composed of identical units and that the system is executing a set of partially ordered tasks with known execution times, using a non-preemptive scheduling strategy.

Ghanem presents two papers. In the first he describes an algorithm used to divide the main memory among competing programs in a multiprogramming environment in a virtual memory system. He uses the working-set concept to determine an optimal allocation policy by developing a variable partitioning algorithm. (The concept is based on an optimization criterion expressed analytically by a function of the working-set sizes and their derivatives.) The optimal allocation is obtained by minimizing the page-fault rates of the programs.

In his second paper, *Ghanem* investigates the effect of the shape of the lifetime function on the optimal partition of the main memory among the programs. Here the optimization criterion is based on maximizing the utilization of the CPU.

Lavenberg and Slutz have also contributed two papers to this issue. In the first, the authors present an introduction to regenerative simulation. The paper is basically tutorial, but it also considers the pragmatic issue of the simulation duration required to obtain valid estimates by a method which is applicable (only) if the stochastic system being simulated is regenerative and if the first two moments of certain random variables associated with a tour are finite.

In their second paper, *Lavenberg and Slutz* introduce new techniques for estimating confidence intervals when stochastic systems with a regenerative structure are being simulated. They describe how these techniques can be applied to the simulation of a queuing model of a computer system's automated tape library. Their model consists of an open network of interconnected queues and incorporates scheduling algorithms, finite capacity queues, and non-exponential service times. The authors conclude that regenerative simulation is a viable tool for the numerical study of a complex queuing model that is not analytically tractable. They describe in this paper in great detail the assumptions made and the restrictions imposed by their model.

Chang presents sequential server queues for the analysis of computer communication systems. The author studies the queuing behavior of sequential server systems with Poisson arrivals, general service time distributions, and several service disciplines, including bulk arrivals, message priorities, and input and output queues. He uses embedded Markov chains for the determination of the stationary distributions of queue lengths and waiting times and illustrates the application of the models to a variety of practical problems with numerical examples.

Wu and Chen have developed a finite population, multi-queue model for a loop transmission system with round-robin scheduling of services. The queues are served in cyclic order by a traveling server and the authors suggest that such a model describes many terminal-oriented computer communications systems. They show that significant parameters, such as average message response time, average cycle time, and average response time conditioned on message length, can be obtained. They have developed several recursive expressions for this model for obtaining the state transition matrix.

Herzog discusses optimal scheduling strategies for real-time computers. He shows how to describe and analyze arbitrary combinations of preemptive and non-preemptive priority strategies and determines the optimal priority strategy by considering the constraints on the response time. He states that preemption-distance priorities with fixed interrupt levels, introduced in this paper, guarantee a fast response time to urgent requests while minimizing software overhead and hardware cost.

Ron Ashany
Associate Editor