Preface

Our motivation for the publication of this tutorial comes from the profound importance and impact of scheduling and load balancing methods on parallel and distributed systems. Concurrent processing in general, and scheduling and load balancing in particular, have been the subjects of research and development during the past two decades. Since the late 1980s, newly available commercial concurrent systems have heightened interest in the areas of scheduling and load balancing. For example, the number of papers published in this area, both in journals and conference proceedings, has increased continually over the past few years. In addition, several workshops and special issues of journals have recently been dedicated to the topic of scheduling and load balancing in parallel systems. Some examples include:

- a special issue of the *Journal of Parallel and Distributed Computing*, on "Scheduling and Load Balancing," co-edited by Behrooz Shirazi and A. R. Hurson, December 1992;
- a workshop on "Scheduling," in the Supercomputing Conference, coordinated by John Feo and Behrooz Shirazi, November 1992;
- a dedicated track on "Scheduling and Load Balancing," in the Hawaii International Conference on Systems Sciences, coordinated by Behrooz Shirazi and A. R. Hurson, January 1993;
- a dedicated track on "Program Partitioning and Scheduling in Parallel and Distributed Systems," in the Hawaii International Conference on Systems Sciences, coordinated by Min-You Wu, January 1994;
- a dedicated track on "Partitioning and Scheduling for Parallel and Distributed Systems," in the Hawaii International Conference on Systems Sciences, coordinated by Apostolos Gerasoulis and Tao Yang, January 1995.

Many of the international conferences in the areas of parallel or distributed processing have allocated several sessions to both static and dynamic scheduling topics in recent years. Examples of such conferences include the IEEE Symposium on Parallel and Distributed Processing, the International Conference on Parallel Processing, the International Parallel Processing Symposium, and the International Conference on Distributed Computing Systems.

Finally, many journals, such as *IEEE Transactions on Parallel and Distributed Systems*, *Journal of Parallel and Distributed Computing*, *IEEE Transactions on Software Engineering*, and *IEEE Parallel and Distributed Technology,* routinely publish papers in the areas of static scheduling and dynamic load balancing.

The central goal of this book is to provide an overview, a detailed discussion, and a prognosis of future directions of the static scheduling and dynamic load balancing methods in parallel and distributed systems. The book covers a wide range of topics, from theoretical background to practical state-of-the-art scheduling and load balancing techniques. However, as scheduling and load balancing can potentially cover a very broad range of topics, we limit our coverage to the following specific subject matters:

- static task scheduling,
- partitioning and task granularity issues,
- scheduling tools,
- load balancing and load sharing,
- task migration, and
- load indices measurement techniques.

It should be noted that there are some important topics, relevant to the subject of this tutorial, that are not covered here, including system (architecture) partitioning, data partitioning, trace scheduling, loop scheduling, user-level threads, and real-time scheduling. It should be noted that real-time scheduling is already covered in another IEEE book (*Tutorial on Hard Real-Time Systems,* by John A. Stankovic and Krithi Ramamritham, 1988.).

This book is intended to be useful to a large number of readers working on different aspects of parallel and distributed systems, including industry professionals, academic professors, and students, who are involved in research or development in the following areas:

- parallel processing applications;
- compilers and operating systems for parallel or distributed systems;
- software tools for parallel program development; and
- system design for parallel and distributed systems, in general.

The readers are expected to have a basic background in the field of parallel or distributed processing. Therefore, the level of the material presented in this tutorial is in the *intermediate to advanced* range.

To the best of our knowledge, there are no books in the market today that specifically address the topic of scheduling and load balancing in parallel and distributed systems. Some closely related books include *Introduction to Parallel Computing*, by Ted G. Lewis and Hesham El-Rewini, Prentice Hall, 1992, *Partitioning and Scheduling Parallel Programs for Multiprocessors*, by Vivek Sarkar, MIT Press, 1989, *Parallel Programming and Compilers*, by Constantine D. Polychronopoulos, Kluwer Academic Publishers, 1988, *Assignment Problems in Parallel and Distributed Computing*, by Shahid Bokhari, Kluwer Academic Publishers, 1987, and *Distributed Operating Systems*, by Andrzej Goscinski, Addison-Wesley, 1992.

Thus, we strongly feel that it is time to publish a comprehensive tutorial to address the important topic of scheduling and load balancing in parallel and distributed systems.

There is a large body of literature on the subject of scheduling and load balancing. This has made the preparation of this manuscript rather difficult since many outstanding and important papers could not be included due to space limitations. To compensate for this problem and provide a more comprehensive coverage of the topic, we have included an extensive bibliography in the introduction section of each chapter. We hope such sections will help interested readers to identify more easily the important papers in their areas of endeavor.

Behrooz A. Shirazi
The University of Texas at Arlington
Ali R. Hurson
Pennsylvania State University
Krishna M. Kavi
The University of Texas at Arlington