

How Robust are Resource Allocations for Tasks in Parallel and Distributed Computing Systems

Prof. H. J. Siegel
Department of Electrical and Computer Engineering
and Department of Computer Science
Colorado State University
E-mail: HJ@ColoState.edu
www.engr.colostate.edu/~hj

Abstract

Performing computing and/or communication tasks on parallel and distributed systems involves the coordinated use of different types of machines, networks, interfaces, and other resources. These resources should be allocated to the tasks in a way that maximizes some system performance measure. However, decisions about how to allocate resources are often based on estimated values of task and system parameters. These estimated values are also used as a basis for predicting the system performance that will result from a given resource allocation. The actual values of these parameters may differ from the estimates for many reasons; for example, the estimates may represent only average values, the models used to generate the estimates may have limited accuracy, and there may be changes in the environment. Thus, an important research problem is the development of resource management strategies that can guarantee a particular system performance given such uncertainties. To address this problem, we have designed a methodology for deriving the degree of robustness of a resource allocation — the maximum amount of collective uncertainty in system parameters within which a user-specified level of system performance (QoS) can be guaranteed. The foundation of this methodology is our mathematical formulation of a metric that evaluates the robustness of a resource allocation. Our four-step procedure for deriving a robustness metric for an arbitrary system will be presented. We will illustrate this procedure by using it to derive robustness metrics for some example distributed systems. Furthermore, we will demonstrate the ability of the robustness metric to select the most robust resource allocation from among those that otherwise perform similarly (based on the primary performance criterion). These methods are applicable to the allocation of resources for many different types of computing and/or communication tasks in many types of environments, including parallel, distributed, cluster, grid, Internet, embedded, wireless, and reconfigurable systems.

This research was performed jointly with Prof. Shoukat Ali, Prof. Anthony A. Maciejewski, and Mr. Jong-Kook Kim. It was supported by the DARPA and by the Colorado State University George T. Abell Endowment.

Speaker

H. J. Siegel was appointed to the endowed chair position of Abell Distinguished Professor of Electrical and Computer Engineering at Colorado State University (CSU) in August 2001, where he is also a Professor of Computer Science. In December 2002, he became the first Director of the CSU Information Science and Technology Center, a university-wide organization for promoting, facilitating, and enhancing CSU's research, education, and outreach activities pertaining to the design and innovative application of computer, communication, and information systems. From 1976 to 2001, he was a Professor in the School of Electrical and Computer Engineering at Purdue University. He received two B.S. degrees from the Massachusetts Institute of Technology (MIT), and the M.A., M.S.E., and Ph.D. degrees from Princeton University. He is a Fellow of the IEEE and a Fellow of the ACM. Prof. Siegel has co-authored over 300 published technical papers in the areas of parallel and distributed computing and communications. He was a Coeditor-in-Chief of the *Journal of Parallel and Distributed Computing*, and was on the Editorial Boards of the *IEEE Transactions on Parallel and Distributed Systems* and the *IEEE Transactions on Computers*.