CC PUB #272

April 1999

Judicious Exploitation of Redundancy for Performance Enhancement of Parallel Resource Systems

Yitzhak Birk and Noam Bloch

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

Abstract

Many parallel resource systems, such as parallel or distributed computing and storage systems as well as networks with multiple paths, exhibit various levels of redundancy. This permits some flexibility in the assignment of tasks to resources, which can be used for load balancing. We propose and analyze assignment schemes that combine elements such as the use of state information, submission of redundant, low-priority requests, error-correcting codes, selective exploitation of redundancy and different binding times. For "code-based" join-the-shortest-queue (JSQ), whereby each task is dispersed into $m \geq 1$ subtasks that are allocated to the m shortest-queue among $m+r$ randomly chosen queues, we show that even a small amount of redundancy suffices for obtaining an asymptotically exponential improvement in the rate of decay of queue-occupancy probabilities over random assignments. To mitigate the effect of imperfect queue-state information, we propose and analyze a hybrid between JSQ and Prioritized Dispersal: m subtasks are assigned a high priority and submitted to the (supposedly) shortest queue(s), and the remaining r subtasks are assigned low priorities and are submitted to other queues. JSQ-PD is shown to substantially outperform JSQ and PD across the reliability range of the state information. Finally, we suggest and briefly discuss a "Flood and Purge" family of schemes, which augment JSQ-PD by purging from the queues subtasks that are no longer required.