The Journal of Supercomputing, 28, 5–6, 2004

Efficient Resource Utilization in Parallel and Distributed Systems

TEOFILO F. GONZALEZ, GUEST EDITOR teo@cs.ucsb.edu

Department of Computer Science, University of California, Santa Barbara, CA, 93106

Abstract. For this special issue we have selected five papers that address, from several points of view, the problem of efficient utilization of resources in parallel and distributed systems. These papers were among the best papers presented at the IASTED PDCS 2001 conference. The topics covered include: efficient cache strategies for simultaneous execution of threads as well as for the distribution of video-on-demand, efficient communication and failure recovery, and run-time support for the automatic parallelization of dynamic structures

Keywords: parallel and distributed systems, cache strategies, efficient communication, failure recovery, automatic parallelization

The Thirteenth IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS'01) was held in Anaheim, CA on August 21–24, 2001. The conference attracted a wide range of papers in many diverse areas of Parallel and Distributed Computing.

For this issue, we selected a representative sample of some of the outstanding papers that were presented at the conference. Even though the five papers are in different areas of parallel and distributed computing, they all deal with the efficient utilization of resources from several points of view. A dynamic cache partitioning mechanism is introduced by Suh et al. for the efficient simultaneous execution of multiple processors/threads. The VoD paper by Lee et al. presents a higher level cache strategy for the efficient distribution of video-on-demand. Dekel and Goft's paper presents a software architecture for end-to-end QoS which guarantees for Web applications efficient recovery under multiple failures. Li et al. introduce a new processor architecture in which collective communication can be performed efficiently without the need of excessive links as in other architectures. Run-time support for the automatic parallelization of pointer-based dynamic data structures is discussed and implemented in the paper by Chan and Abdelrahman. In what follows we discuss the papers individually.

- "Dynamic Partitioning of Shared Cache Memory" by G. E. Suh, L. Rudolph and S. Devadas.
  The paper proposes dynamic cache partitioning amongst simultaneously executing processes/threads. The general partitioning scheme can be applied to set-associative caches. The idea is to collect information about cache miss characteristics at run-time and then partitions the cache among the executing processes. The partition sizes are varied dynamically for efficiency purposes.

- "An Efficient Scheduling Algorithm for Information Delivery on VoD System" by SingLing Lee, Hang-Jang Ho and Wen-Wei Mai.
  Minimizing transmission and storage cost for delivering programs is a fundamental problem in video-on-demand (VoD) systems. Since the amount of data to be transmitted is quite large, efficient "cache" sites for future delivery of services enhance significantly the delivery of VoD. This work is aimed at VoD systems under a metropolitan-area network.
- "ITRA: Inter-Tier Relationship Architecture for End-to-end QoS" by Eliezer Dekel and Gera Goft.
  The ITRA (Inter-Tier Relationship Architecture) is presented for end-to-end quality of service (QoS). The mechanisms, roles of the tiers, programming model and inter-tier relationship protocol are discussed. Web applications following the ITRA architecture can collaborate to transparently recover from failures in multiple tiers, as well as to better exploit mutual resources to provide the required availability and failover transparency aspects of an end-to-end QoS.
- Efficient Collective Communications in Dual-Cube" by Yamin Li, Shietung Peng and Wanming Chu.
  The dual-cube architecture is introduced for large-scale parallel computation and algorithms for efficient collective communications are described. The dual-cube network retains the hypercube's topological properties without having the large number of links required by large-scale hypercubes. The paper establishes that collective communications can be carried out in the dual-cube with almost the same communication complexity as in the hypercube.
- "Run-Time Support for the Automatic Parallelization of Java Programs" by Bryan Chan and Tarek S. Abdelrahman.
  The paper describes and evaluates a novel approach for the automatic parallelization of programs for pointer-based dynamic data structures. The approach is to exploit parallelism among methods by creating an asynchronous thread of execution for each method invocation in the program. At compile time information about the methods is collected and analyzed. The run-time system, which is the main focus of the paper, uses this information execute correctly and efficiently the code. The approach is validated empirically.

I hope you find all of this work useful, and that you consider contributing to the IASTED PDCS conference in the near future. The conference provides a friendly environment where all aspects of parallel and distributed computing are discussed.