中央研究院 資訊科學研究所
Institute of Information Science, Academia Sinica

Distinguished Lecture Series

Warehouse-scale Computers: Opportunities and Challenges

Monday, May 5th, 2014 10:00am
Auditorium 106 at New IIS Building

Mary Lou Soffa

Department of Computer Science, University of Virginia

Abstract

Web-service companies such as Google, Microsoft, Amazon, Yahoo, and Apple spend hundreds of millions of dollars to construct and operate Warehouse-scale Computers (WSC) which provide popular web-services such as search, social networking, webmail, video streaming, enterprise management tools, online maps, automatic translation, and online courses. The primary advantages of WSC are the scalability and cost benefits for both the end-users and web-service companies. These WSCs house hundreds to thousands of machines to provide the computing resources needed to serve millions of users. To limit the cost of ownership of WSCs, these machines are composed of commodity components which are cheap and easily replaceable, often multi-cores. When multiple applications are running simultaneously on a multi-core machine, resources sharing and contention among cores can result in a significant amount of performance interference. This interference leads to a significant problem in meeting the requirements of user facing web-service applications. To avoid the constant unpredictable threat that shared resource contention poses to an application's QoS, datacenter operators typically disallow co-locations of latency-sensitive jobs with other jobs. This unnecessary over-provisioning of computer resources reduces the overall utilization of WSCs and results in an unnecessarily high cost and a large environmental footprint for a given set of web-service workloads. In this talk, I discuss these issues and present our research using scheduling and compiling to improve the capability and cost effectiveness by improving resource efficiency. Specifically, we reconcile the apparent conflict between the need to maintain high QoS for latency-critical, high-priority services and the desire to increase hardware utilization by scheduling multiple workloads per server.

For more infomation: http://www.iis.sinica.edu.tw/