

ECE/CS 670 - Topics in Architecture/Systems - (C) Distributed Systems

Heterogeneous Computing

Electrical and Computer Engineering Dept.
and Computer Science Dept.
Colorado State University
October 30, 2009

Prerequisite: Any one of: ECE 550, ECE 554, ECE 561, CS 551, CS 570,
CS 575, or consent of instructor.
Course Credits: 4

Instructor: Professor H. J. Siegel
Office: Engineering B115
Phone: 491-7982
Email: HJ@ColoState.edu

Description

In a heterogeneous computing environment, a suite of different machines is interconnected to provide a variety of computational capabilities to execute collections of application tasks that have diverse requirements. The execution times of a task will vary from one machine to the next, and tasks will compete for machines in the suite. There are many types of heterogeneous systems, including parallel, distributed, clusters, and grids. They can be found in industrial, laboratory, government, academic, and military settings. Such systems may be used in production, computing center, embedded, or real-time environments. An important research problem for heterogeneous computing is how to assign computation and communication resources to tasks and to schedule the order of their execution to maximize some performance criterion, a process known as mapping or resource management. Factors that must be considered include machine and network loading, how well the execution needs of a task match the computational capabilities of a machine, any inter-task communications, operating constraints, and the performance criterion to be optimized.

The class lectures will include the following material. An overview of the field of heterogeneous computing will be given. Dynamic and static heuristics for mapping tasks to resources in a heterogeneous system will be presented. Tasks that involve priorities, deadlines, and alternate versions of different worths to the user will be considered. The design of resource allocations that are robust against uncertainties will be studied. Open problems in the field of heterogeneous computing will be discussed.

This course is intended for ECE, CS, and other graduate students who want to learn about the ways in which a collection of heterogeneous machines can be used to execute a single large application task, a set of independent tasks, or sets of interrelated application tasks in a way that will optimize some performance criterion. The

material is applicable to various types of heterogeneous computing and communication environments, including parallel, distributed, cluster, grid, Internet, cloud, embedded, multicore, content distribution networks, wireless networks, and sensor networks. Furthermore, the resource allocation techniques and robustness models, concepts, and metrics presented are generally applicable to design problems throughout various scientific and engineering fields.

In addition to two 75-minute lectures each week, there will be a one-hour recitation. The recitations will focus on the projects for the course. During the recitations, students will meet with several faculty (Professors Siegel, Pasricha, Potter, and Smith) to discuss research papers that pertain to the application domain of the projects. The knowledge about resource allocation for heterogeneous systems in the lectures will be applied to this domain for the two projects. As in all previous years that this course has been taught, the projects will be published as conference and journal papers, coauthored by all of the students in the class.

The projects for Spring 2010 will focus on robust resource allocation in warehouse scale heterogeneous computing systems, such as those used by Google, Amazon, and Yahoo. These environments involve thousands of heterogeneous machines, where each machine could contain hundreds of heterogeneous cores. Each machine may contain cores that operate in SIMD parallelism (data parallel and associative computing) or in MIMD parallelism (similar to a multiprocessor). Collections of these processors could be operating on a few huge data sets, or a huge number of smaller data sets. Background technical papers and the exact robust resource allocation problems to be addressed in the projects will be discussed in the recitation classes. Related topics, such as operating systems, programming languages, and compilers as they apply to warehouse scale computing, also will be discussed.

Course Objectives

This course will enable the student to:

- understand the potential advantages of using heterogeneous computing systems
- analyze some of the factors that must be considered when designing resource management systems for heterogeneous environments
- be familiar with a variety of *dynamic* (on-line, real-time) and *static* (off-line) techniques for assigning resources to tasks and scheduling their execution to optimize some performance criterion
- evaluate the effectiveness of resource allocation schemes for heterogeneous systems
- formulate a performance metric appropriate for the goals of a given computing environment
- build robustness models and metrics for analyzing resource allocations and designing resource management heuristics that must cope with uncertainties in system and workload parameters
- be aware of the open research problems in heterogeneous computing that are important areas for future research and development.

Course Materials

This course will be taught using technical papers from the literature.

There will be evening exams.

Course Policies (subject to change)

1. Projects

There will be two research projects, both of which may involve programming and simulation studies. Students will work in pairs on these projects.

2. Exams

- a. There will be three exams, each covering approximately one-third of the course; there will be no comprehensive final (i.e., no final exam that covers the entire course).
- b. Two exams will be given in the evening; the third exam will be given in the final exam time slot.

3. Grading

- a. Your final course grade will be based on your projects and examinations. Each of the two projects will be worth 20%, and each of the three exams will be worth 20%.
- b. Your letter course grade will be determined from the total points that you obtain from your projects and tests, and will be based on a combination of a relative and an absolute scale. You determine your own grade by your performance on these items.

Course Outline for Lectures (Tuesday/Thursday 11 to 12:15 p.m.)

Topic	Weeks
1. Class policy, introduction to heterogeneous computing, automatic heterogeneous computing, open problems	1
2. Static mapping of applications composed of communicating tasks	1
3. Static mapping techniques for independent tasks	1
4. Dynamic mapping techniques for independent tasks	1
5. Exam 1	0.5

6.	Robust resource allocation: <u>deterministic</u> model and heuristics	2
7.	Robust resource allocation: <u>stochastic</u> model and heuristics	2
8.	Mapping of tasks with priorities, deadlines, and versions	1
9.	Exam 2	0.5
10.	Student presentations of first project	1
11.	Scheduling communications in overloaded networks	1
12.	On-line use of off-line derived mappings	0.5
13.	Advanced current topics	1.5
14.	Student presentations of second project	1

Course Outline for Recitations
(at a time to be arranged based on students' schedules)

Topic	Weeks
1. Introduction to warehouse scale computing	2
2. Description and discussion of first project	3
3. Advanced aspects of warehouse computing, such as reliability and power considerations	2
4. Description and discussion of second project	3
5. Student presentations of first project	1
6. Related topics, such as operating systems, programming languages, and compilers as they apply to warehouse scale computing	3
7. Student presentations of second project	1