CIS 750. High Performance Computing : Web Searching

Course Description

An in-depth study of the state of the art in high performance computing with **the emphasis this semester being on high-performance Web-searching techniques**. Topics include parallel computer architectures, programming paradigms, the Google File System, the Google MapReduce model, Web-crawling, Web search and high-performance information retrieval, and their applications. Parallel programming paradigms include the message passing interface (MPI), and its second-generation MPI-2. Applications include computational science and high-performance Web searching. First-hand experience in stable, scalable, high performance computing for Internet-based application design.

Contact Information

Instructor: Alex Gerbessiotis Lectures : Monday 6-9pm E-mail: alg750@cs.njit.edu

Web : http://www.cs.njit.edu/~alexg/courses/cis750/index.html

Prerequisites

CIS 650. Why?	It's in the catalog.
Real Prerequisites	Students are expected to complete a number of programming assignments related to the construction of a high-performance parallel search engine. The high-performance portion (one of four) of the programming assignments requires knowledge of C++ or C (eg. pointers) and will use MPI-2. However the rest of the assignments can be done in C++, or C, Java, Perl or any other language combination, as long as this combination is available on the testing platform/PC cluster. Students can also use standard Unix packages (eg. Flex/Lex etc).
HTML knowledge	The basics of HTML; you will search HTML documents, so you need to know what they look like. If you don't know HTML, then you can download information from the Web and learn the basics in an hour or so.
Logistics	Students will be given access to the testing platfrom (a PC cluster) or they can also use their own PCs and/or AFS accounts.
Textbook	Modern Information Retrieval by R. Baeza-Yates and R. Ribeiro-Neto, Addison Wesley.
Other books	Notes (Scribing). Papers (links to be given). MPI-2 (notes to be handed out) or : Using MPI - 2nd Edition: Portable Parallel Programming with the Message Passing Interface (Scientific and Engineering Computation) by William Gropp, Ewing Lusk, Anthony Skjellum. MIT Press; 2nd edition (November 26, 1999), ISBN: 0262571323.
CourseWork:	A group project in four parts (max group participation : 2) with final-report (50%), homeworks in the form of one-page paper summaries (25%), individual Work (project extension or literature review) (25 %).
Group Project:	*A simple high-performance parallel search engine.*
Individual Work:	*An individual extension of the Group project (it can reuse group-code) or a literature-review of the student's choice (10-page, 12 point, PDF).* In the latter case, a 15-minute presentation (plus 5 minutes for questions) will take place in one of the last three weeks of classes. Slot assignments first come first served.
Enrollment:	Limited to 18.