New Jersey Institute of Technology

A. V. Gerbessiotis | CIS 750
Jan 20, 2006 | Spring 2006
Course Syllabus: General Information | Handout 1

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, parallel programming paradigms (eg MPI and MPI-2), the *Google File System*, the *Google MapReduce model*, Web search and high-performance information retrieval, and their applications. First-hand experience in stable, scalable, high performance computing for Internet-based application design.

Contact Information

Instructor:	Alex Gerbessiotis	E-mail:	`alg750@cs.njit.edu`
Office:	GITC 4213, 4th floor	Tel:	(973)-596-3244
Office Hours:	Tue 11:30-1pm, Mon and Tue 4:30-5:30pm		
Office Hours:	By appointment some other time on Mon,Tue,Wed		
Class Hours:	Mon 6-9pm, Tiernan B8		
Course Web Page:	`http://www.cs.njit.edu/~alexg/courses/cis750/index.html`		

Print this Handout 1 from Web-page and compare the printout to this document! They must be identical.

Course Administration

Prerequisites CIS 650.

Textbook Modern Information Retrieval by R. Baeza-Yates and R. Ribeiro-Neto, Addison Wesley, ISBN 0-201-39829-X.

Recommended But not required: *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.

Other Papers will also be used and lecture notes summaries will become available. Links to papers or local copies can be found in the protected area of course the web-page.

CourseWork
1. A group project in four parts. Max Group participation is 2.
2. A final project report.
3. Homeworks in the form of one-page paper summaries.
4. Additional individual work (e.g. project extension, or literature review with presentation. If individual work is a project extension it CANNOT be a group collaboration.

Grading scheme 1000 points maximum.

1. Project : 450 points
2. Project report (full documentation) : 50 points
3. Paper summaries : 250 points. Each summary is worth around 40-60 points. A student can do as few or as many as he/she likes but she/he can collect no more than 250 points.
4. Project Extension/Paper Review: 250 points (of which 70 points to be determined by the in-class presentation). Student decides the topic. Presentation in one of the last two weeks of classes.

Due Dates Programs **MUST be received by email before midnight the day** they are due. Other assignments are due by start of class but no later than 6:05pm. Late-submission penalties apply otherwise (see Course Policies, page 3).

Tentative Course Calendar

Spring 2006					
Week	Mon	PSout	PSin	PAin/out	Comments
W1	1/23	PS1out			
W2	1/30		PS1in	PA1out	Groups-In
W3	2/6	PS2out			
W4	2/13			PA2out	
W5	2/20	PS3out	PS2in		
W6	2/27			PA1in, PA3out	
W7	3/6		PS3in	PA4out	Proposal for Ind. Work due
W-	3/13	Break	Break	Break	Spring Break
W8	3/20	PS4out		PA2in	
W9	3/27	PS5out	PS4in		
W10	4/3	PS6out	PS5in	PA3in	
W11	4/10	PS7out	PS6in		
W12	4/17		PS7in		
W13	4/24	Presentation I		PA4in	
W14	5/1	Presentation II		Project report/Individual work due	
W14	5/3	No CLASSES		No CLASSES	Reading Day:Absolute Deadline
W15	5/8	———	———	Exam Period	

PA= Programming Assignment (Group Project). PS = Paper Summary (done individually).
*. Keywords with stars next to them indicate a paper will be handed out for further reading/discussion.

```
T1 : Introduction. Searching the Web. The process. Robots/Crawlers.            [Ch 1, 13]
     Measuring the Web. Modeling the Web*. Search Engines. Ranking.
     Indices. Metasearches.
T2 : Text Markup Languages. SGML.HTML.XML. Zipf's Laws.                         [Ch 6]
T3 : Text Operations. Doc Processing. Stopwords. Stemming. Index term           [Ch 7,8]
     selection. Compression. Huffman coding. Inverted files.
     Inverted file compression. Indexing. Searching. Signature Files.
     Searching with errors.
T4 : Modeling. Taxonomy of information retrieval Models                         [Ch 2, 3, 4]
     (Boolean, Vector, Probabilistic).  Models for browsing.
     Retrieval Evaluation (Recall and Precision)
     Alternative measures. Query Languages
T5 : The Google Cluster Architecture*. The Google File System*.
     PageRank*.
T6 : Introduction to Parallel Computing. Flynn's Taxonomy.                      [Ch 9]
     Amdahl's Law. Gustaffson's Law; Brent's Principle. PRAMs.
     Types of PRAMS (EREW, CRCW).
T7 : Fundamental Operations on PRAM.
     Parallel Min/Max and parallel Sum. Parallel Prefix Computations.
     Parallel Addition. Broadcasting. Matrix multiplication.
T8 : Parallel Algorithm Design on the BSP model.Introduction.
     Using RMA (Remote Memory Access) for parallel programming.
T9: Parallel Programming: LAM-MPI and MPI-2.
T10: Parallel Information Retrieval and Inverted Files.                         [Ch 9]
T11: Parallel Web-crawlers*. The Google Map-Reduce Model*.
T12: Parallel Analysis with Sawzall*. Extracting knowledge from the World Wide Web*.
```

Any modifications or deviations from these dates, will be done in consultation with the attending students and will be posted on the course Web-page. It is imperative that students check the Course Web-page regularly and frequently.

Written Work Can be handwritten or typewritten. In the former case, make sure that a person other than you can read what your wrote. DO NOT USE pencils to write down your answers; if you decide to use a pencil do not complain about grading.

Programs Code must be ANSI compliant and compilable on the test platform/compiler. Follow the guidelines provided for input/output interfaces.

Grading Written work will be graded for conciseness and correctness. Use formal arguments, if required. Be brief and to the point. Do not repeat the sentences of the paper you are summarizing. Programming problems will be graded based on test instances decided by the grader on a test platform of his choice. Do not expect partial credit if your code fails to run on all test instances. Do not expect any partial credit if your code does not compile. Excess Programming points related to implementations beyond that outlined in the assignments can gain you extra points.

Extension policies Each student is given a seven-day grace period for late work. Minimum delay unit is a day. For example a non programming assignment received at 6:06pm, i.e. one-minute late according to the instructor's watch, not YOURS!, uses up one day; a programming assignment received at midnight or as late as 23:59 the following day also uses one day. Late programming assignments use up days of both members of a group. When a student hits the 8-th day of extensions (one day beyond the grace period of 7 days) his grade will be lowered by one half-level (an A will become a B+); a 13-th day, 18-th, and 23-rd day also trigger similar penalties. **No extension for the final project report. No piece of work will be accepter after the Reading Day.**

Grade questions Check the marks in a written work and report errors promptly. **Make sure you report such problems to the grader or the instructor within two weeks from receipt** but **no later than the Reading Day**. If you believe a grade you received for the solution of a problem is not representative of your effort talk to the grader first and then to the instructor (if different).

Final Grade The final grade is decided based on the 0 to 1000 point performance with an adjustment made based on programming assignment performance. A student who collects at least 500 points should expect a passing grade (C). The instructor reserves the right to push a student's grade up based on that student's quality of his/her programming effort.

Collaboration Only the group project is a collaborative work of the members of the group. Every other coursework must be done individually. Students who turn in solutions (programming or otherwise) that are derived from solution outlines of past assignments/homeworks, were obtained through the Internet, or are a product of another student's work, risk severe punishment, as outlined by the University. The work you turn in MUST BE your own personal work, composed and written by you. If you talk a problem with a fellow student cite this clearly in your homework (name the fellow student before the solution of the problem in question). Your work will then be compared to the other student's work to verify that your solution was written by you and reflect your own personal effort. If you don't report it, it will be considered a violation of the course rules.

Mobile Devices Switch off noisy devices (e.g. mobile phones) before you enter the classroom for a lecture.

Email/SPAM Send email from an NJIT email address. NJIT spam filters or us will filter other email address origins. Do not send course email to the instructor's email address unless there is a good reason (e.g. you didnot get a prompt response when you sent an email to the course email address and you suspect email problems) Include `CIS 750` in the subject line then. ■.

The NJIT Honor Code will be upheld; any violations will be brought to the immediate attention of the Dean of Students. Read this handout carefully!