IS698/MATH700/PHYS650: Special topics on Big Data + High Performance Computing + Atmospheric Sciences

Instructors

- Dr. Jianwu Wang (jianwu@umbc.edu), Department of Information Systems, UMBC
- Dr. Matthias K. Gobbert (gobbert@umbc.edu), Department of Mathematics and Statistics, UMBC
- Dr. Zhibo Zhang (zhibo.zhang@umbc.edu), Department of Physics, UMBC
- Dr. Aryya Gangopadhyay (gangopad@umbc.edu), Department of Information Systems, UMBC

TAs

- Pei Guo (peiguo1@umbc.edu), focusing on Big Data
- Carlos Barajas (barajasc@umbc.edu), focusing on High Performance Computing
- Chamara Rajapakshe (charaj1@umbc.edu), focusing on Atmospheric Sciences

Course Description

This graduate-level course for students in three disciplines (Computing, Mathematics, and Physics) to foster multidisciplinary research and education using advanced cyberinfrastructure (CI) resources and techniques. The course will teach students how to apply knowledge and skills of high-performance computing (HPC) and Big Data to solve challenges in Atmospheric Sciences. We focus on the application area of atmospheric physics and within it radiative transfer in clouds and global climate modeling, since these topics are important, pose computational challenges, and offer opportunities for big data techniques to demonstrate their impacts.

Instructional Methods

The participants in the new initiative will be selected competitively to form multidisciplinary teams of three participants with one participant from each area. The material is at the level of an advanced graduate course. All work is conducted in an multidisciplinary team with participants from each area, mentored by a faculty and supported by a teaching assistants (TA) from each area. In the first 10 modules consisting of instruction in all three areas, team building is achieved by homework. In the final 5 modules, each team applies the material learned immediately to a small research project, culminating in a technical report and a project presentation.

Grading policy

- Homework: 44%, 4 point for each homework including Homework 0
- Participation: 11%
- Project: 45%

Homework policy: Each homework will be done by the home team except homework 0. Each homework is due on Thursday before the next class.

Communication: Discussions at Blackboard for homework and emails for projects.

Academic Integrity

By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC's scholarly community in which everyone's academic work and behavior are held to the highest standards of honesty. Cheating, fabrication, plagiarism, and helping others to commit these acts are all forms of academic dishonesty, and they are wrong. Academic misconduct could result in disciplinary action that may include, but is not limited to, suspension or dismissal. See more at Academic Policies/Student Rights and Responsibilities at UMBC Gradudate School.

Schedule

Module	Date	Topic	Goal	Instructor
0	1/25	Online communication testing and introduction	Know instructors, TAs and team members	All
1	2/1	Introduction of Python/C, Linux and HPC environment	Running their own jobs on HPC	Gobbert
2	2/8	Numerical methods for Partial Differential Equations (PDE)	Model as PDE and solve them using numerical methods	Gobbert
3	2/15	Message passing interface (MPI)	Write MPI jobs	Gobbert
4	2/22	Basics of earth-atmosphere radiative energy balance and global warming	Understand basic concepts and principles of radiative energy balance and global warming	Zhang
5	3/1	Basics of radiative transfer simulation framework	Understand the basic physics underlying the transport of radiation in atmosphere	Zhang
6	3/8	GCM simulation and satellite observations	Understand the importance of GCM and satellite remote sensing	Zhang
7	3/15	Basics of data science and machine learning	Understand the basics of data science and different types of machine learning tasks.	Gangopadhyay

	3/22	No meeting. UMBC Spring Break.		
8	3/29	Introduction of and Big Data	Understand the basics of Big Data and demo programs	Wang
9	4/5	Big Data system: Hadoop/Spark	Write Hadoop/Spark jobs and run them on HPC	Wang
10	4/12	Big Data Machine learning	Write a machine learning program using Spark MLlib	Wang
11	4/19	Project introduction and assignment	20 minutes project explanation from each team, including Q&A	All
12	4/26	Project progress report from each team and feedback from instructors	20 minutes report from each team including Q&A + rating	All
13	5/4	Project progress report from each team and feedback from instructors	20 minutes report from each team + Q&A + rating	All
14	5/10	Project progress report from each team and feedback from instructors	20 minutes report from each team + Q&A + rating	All
15	5/17	Final project presentation	Technical report, software and a final 30 minutes presentation from each team (by all team members) including Q&A.	All

Module 1: Introduction of Python/C, Linux and HPC environment. The first module explains the whole structure of the program and required basic knowledge for the program. It briefly goes through a programming language such as Python or C. It also introduces the hardware architecture, available software and basic usage of the UMBC HPCF environment.

Module 2: Numerical methods for Partial Differential Equations. This module will explain the basics of partial differential equation, which is commonly used in physical models. It will discuss the use of numerical methods for PDEs, which is one major driving force behind research in many other fields like numerical linear algebra, scientific computing, and the development of parallel computers. It will cover the three basic PDE categories and their mathematical properties with examples. It will discuss two large classes of methods: finite difference and finite element methods.

Module 3: Message Passing Interface (MPI). This module will explain how to write MPI programs which is one of most common approach to build portable and scalable parallel scientific applications. It will cover basic MPI commands such as MPI_Send and MPI_Recv,

collective communication commands like MPI_Bcast, MPI_Reduce/MPI_Allreduce, and MPI_Gather/MPI_Scatter. It will also explain how to write MPI programs in both C and Python (through mpi4py).

Module 4: Basics of earth-atmosphere radiative energy balance and global warming. This module will explain the basic concepts and principles that control the radiative energy balance of earth-atmosphere system, and its implications to climate. The module will start with the fundamental physics, such as black-body radiation, followed by zero-order radiative energy balance between incoming solar radiation and outgoing terrestrial longwave radiation. The module will end with discussion of what kinds of roles the greenhouse gases, aerosols and clouds play in the radiative energy budget.

Module 5: Basics of radiative transfer simulation framework. Following previous module, this module will introduce the fundamental physical principles that control the transport of radiation (i.e., visible and infrared light) in our atmosphere. The module will also include the introduction of Monte-Carlo method and it application to radiative transfer.

Module 6: GCM simulation and satellite observations. This module will start with an introduction to the basic concepts and principles of numerical climate simulations, followed by explaining the importance of evaluating climate simulations and why satellite remote sensing products are invaluable for climate model evaluation. Basic concepts and principle underlying satellite remote sensing will also be introduced this module.

Module 7: Basics of data science and machine learning. This module will first explain the basic concepts of Data Science, including generic lifecycle and different stages of data analytics, such as acquisition, cleaning/preprocessing, integration/aggregation, analysis/modeling and interpretation. This module will then explain the main lifecycle (training, testing, applying) and main types of machine learning (supervised and unsupervised learning). Major techniques to be covered include inferential statistics, feature selection, regression, correlation, clustering, classification and anomaly detection.

Module 8: Introduction of Big Data. This module will explain the basic concepts of Data Science, including generic lifecycle and different stages of data analytics, such as acquisition, cleaning/preprocessing, integration/aggregation, analysis/modeling and interpretation. It will explain the basics of Big Data, including its 5V characteristics. It starts with the challenges and bottleneck of many applications when dealing with large volume of data. It will cover unique features and challenges for satellite data.

Module 9: Big Data system: Hadoop/Spark. This module will cover Big Data concepts/techniques: distributed file system, data partitioning, data parallelization, key-value pairs, functional programming and MapReduce. This module will cover how to use two popular Big Data systems namely Hadoop and Spark. It will explain how Hadoop Distributed File System (HDFS) can achieve data partitioning, and fault tolerance and cluster management and job scheduling in Hadoop/Spark. For Spark, it will explain resilient distributed datasets (RDD), RDD transformations (map, join, cogroup, etc.) and actions (count, collection, foreach, etc.), lazy evaluation.

Module 10: Big Data Machine learning. This module will explain how to conduct machine learning tasks in the above module in a scalable approach through Spark MLlib. Techniques/concepts include DataFrame-based MLlib API vs RDD-based MLlib API, ML pipelines, Transformer, Estimator and Parameter.

Module 11: Project introduction. This module will explain available research projects to be conducted in the following five weeks. For each project, it will cover the required techniques, suggested phases and major tasks, expected outputs, output evaluation metrics and challenges

to each discipline. The projects will be assigned to teams by their mentors ahead of time. During this week, each team will explain the assigned project to all participants and mentors.

Module 12-14: Project progress report from each team and feedback from instructors. These three modules will be weekly project progress updates and discussions. Since each team has three members, every member will be a presenter for the reports. All instructors and other teams will discuss the progress, perform peer review, provide feedback and give ratings.

Module 15: Final project presentation. The final module will be the final project presentation and final CI software program and technical report delivery. Each team will give a talk on the problems to be solved, the approaches taken, demonstration of developed software program, the experiments and results, and contributions of each member. All instructors and other teams will provide feedback and give ratings and suggestions for future work.