ECE 445/519 Recent Advancements in High-Performance Computing (HPC)

COURSE INSTRUCTOR AND CONTACT INFORMATION

- **Name**: Hang Liu
- **Email**: Hang.Liu@rutgers.edu
- **Phone**: (848) 445-0876
- **Office**: CoRE 509
- **Website**: https://www.ece.rutgers.edu/hang-liu

COURSE MEETING DAYS, TIMES, LOCATION, MODALITY

Course time: 12:10 PM – 3:10 PM, Wednesday
Course location: PH-111 (click here [hyperlink] to see the location) of Busch campus
Canvas sites:

- Undergraduate students: https://rutgers.instructure.com/courses/243707
- Graduate students: https://rutgers.instructure.com/courses/245749

Course format: In-person class. See attendance and participation for more policies.

OFFICE HOURS / STUDENT SUPPORT HOURS

Office hours: 9 AM – 12 PM Wednesday, or by appointment via email.
Location (either one below would work):

- Virtual office hour on zoom: https://rutgers.zoom.us/my/hl1097?pwd=RVNmY0ZxVXRmWjd6QzREaGZoZkVqdz09, passcode: hl1097
- In person: CoRE 509, Busch campus

COURSE DESCRIPTION

High-Performance Computing (HPC) is the ability to perform computations at high speeds. While a desktop that runs at 1 GHz can process calculations of around 1 billion per second, which is already much faster than any human being can achieve, today's supercomputer can process beyond 1 billion billion operations per second. Such remarkable processing speeds have made many game-changing innovations possible and improved the quality of life for billions of people around the globe, e.g., HPC has fueled the groundbreaking revolutions for deep learning. In general, HPC is the foundation for scientific, industrial, and societal advancements. In this course, the instructor will provide the students with knowledge about the recent advances in HPC. We will use a three-pronged approach to teaching this course. First, the instructor will give a presentation to the students about how to read papers, discover new ideas, implement them, write technical papers and present them at international venues. Second, the instructor will create a list of well-known HPC papers for the students to choose as their research projects. Note the students are welcome to bring their own projects to this course. Third, the students must finish their projects via proposal, midterm, and final report/presentations. The instructor will encourage and help the students to publish their findings on international venues.

Prerequisites: C/C++/Java programming skills

REQUIRED TEXTS AND COURSE MATERIALS

Reading materials selected by the instructor.

TECHNICAL / TECHNOLOGY REQUIREMENTS

Computers are required to attend this course to complete the assignments and projects.

LEARNING GOALS

After successfully completing this course, the students will be able to:

- Read and summarize a technical paper from the domain of high-performance computing.
- Design and implement a system that is reflected in a selected technical paper.
- Write technical reports for the implemented system.
- Present the design and implementation of the system to a group of peers.

GRADING SCALE

We will follow the default grading scale from Canvas (https://rutgers.instructure.com/courses/243707 and https://rutgers.instructure.com/courses/245749)

ASSESSMENT / GRADING COMPONENTS

Weighting of Assessments

Participation: 10%
The students should attend the course and participate in the discussion of the papers presented by their peers.

Homework 15%
We will read about three papers each week. **Each student is required to read ALL of them ahead of the class and write a short summary for one paper each week.** A template of the paper summary is provided in the Canvas week 1 module. (Please read my Google slides

https://docs.google.com/presentation/d/1fWNPS5Ts2s13T3GFQIo_6H2LEOT1ZVHy-Pu_tDHHp_A/edit#slide=id.p about how to review a paper)

Paper presentation: 20%
Each student would be expected to present and lead the discussion for 1 – 2 papers. **The students must email his/her presentation slides to the instructor (hang.liu@rutgers.edu) 1 week ahead of his/her presentation.** Late or no submissions will result in penalties. (see my Google slides about how to present a paper: https://docs.google.com/presentation/d/1fWNPS5Ts2s13T3GFQIo_6H2LEOT1ZVHy-Pu_tDHHp_A/edit#slide=id.p)

Course project: 25%

- Initial proposal (Week 3)
- Final proposal 5% (Week 4)
- Mid-term report 20% (Week 8)
- Final report 35% (Week 15)

Projects will be done individually or in groups of two students. Each group is required to submit a project proposal, a midterm report, and a final report. **Students are strongly suggested to talk to the instructor often and seek for help as soon as possible.** Each group will present their projects in class. It is expected that publishable results will come out of some projects.

Late submission policy: All assignments are due on Canvas by class 12:10 PM Wednesday. Of note, 20% of the grade will be deducted for each day the assignment is late.

TENTATIVE SCHEDULE OF TOPICS

ATTENTION: We expect 3 paper presentations per lecture. During the lecture, each paper presentation will take ~30 – 40 minutes. The discussion during/after the presentation will take ~15 – 25 minutes. We will have a break of ~5 – 10 minutes after each paper discussion.

TIPS: You can expect the presentation speed to be 1 minute per slide.

WEEK	DATE	TOPIC	CANDIDATE PAPERS (3 PAPER DISCUSSIONS PER LECTURE)
1	9/6	Logistic business	**Choose your papers TODAY!** discuss syllabus, and research tips https://github.com/asherliu/researchHOWTO
2	9/13	Deep learning	1. TensorFlow: A System for Large-Scale Machine Learning [OSDI '16] 2. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning [OSDI '18] 3. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM [SC '21]

3	**9/20**	Machine learning	1. XGBoost: A Scalable Tree Boosting System [KDD ‘16] 2. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks [ArXiv] 3. Scaling Distributed Machine Learning with the Parameter Server [OSDI ‘14]
4	**9/27**	Graph analytics	1. C-SAW: A Framework for Graph Sampling and Random Walk on GPUs [SC ‘20] 2. Everything you always wanted to know about multicore graph processing but were afraid to ask [USENIX ATC '17] 3. CECI: Compact Embedding Cluster Index for Scalable Subgraph Matching [SIGMOD ‘19]
5	**10/4**	Project proposal presentation	15 mins per presentation.
6	**10/11**	Data mining	1. Dr. Top-k: delegate-centric Top-k on GPUs [SC ‘21] 2. Yinyang K-Means: A Drop-In Replacement of the Classic K-Means with Consistent Speedup [ICML ‘15] 3. Billion-scale similarity search with GPUs [IEEE Transaction on BigData]
7	**10/18**	Algorithms	1. Scalable multi-GPU 3-D FFT for TSUBAME 2.0 Supercomputer [SC ‘12] 2. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs [TPAMI] 3. Efficient C4.5 Algorithm [TKDE]
8	**10/25**	Cloud computing	1. FaasCache: Keeping Serverless Computing Alive with Greedy-Dual Caching [ASPLOS ‘21] 2. SkyPilot: An Intercloud Broker for Sky Computing [NSDI ‘23] 3. Firecracker: Lightweight Virtualization for Serverless Applications [NSDI ‘20] 4. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads [NSDI ‘17] 5. From Laptop to Lambda: Outsourcing Everyday Jobs to Thousands of Transient Functional Containers [USENIX ATC ‘19] 6. Shuffling, Fast and Slow: Scalable Analytics on Serverless Infrastructure [NSDI ‘19]
9	**11/1**	Midterm project presentation	15 – 20 mins per presentation.
10	**11/8**	Quantum computing	1. QX: A High-Performance Quantum Computer Simulation Platform [DATE ‘17] 2. QuEST and High Performance Simulation of Quantum Computers [Nature Scientific Report] 3. 0.5 Petabyte Simulation of a 45-Qubit Quantum Circuit [SC ‘17]

11	**11/15**	NO CLASS	Mandatory group-wise project progress discussion
12	**11/22**	NO CLASS	Friday schedule
13	**11/29**	Linear algebra	1. Strassen's Algorithm Reloaded [SC '16] 2. Merge-based sparse matrix-vector multiplication (SpMV) using the CSR storage format [SC '16] 3. Tango: rethinking quantization for graph neural network training on GPUs [SC '23]
14	**12/6**	Applications	1. Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning [SC '20] 2. Highly accurate protein structure prediction with AlphaFold [Nature] 3. Communication-Efficient Jaccard similarity for High-Performance Distributed Genome Comparisons [IPDPS '20] 4. A 400 trillion-grid Vlasov simulation on Fugaku supercomputer: large-scale distribution of cosmic relic neutrinos in a six-dimensional phase space [SC '21] 5. Anton 3: twenty microseconds of molecular dynamics simulation before lunch [SC '21]
15	**12/13**	Final project presentation	20 mins per presentation.

POLICIES

Attendance and Participation

Per GRADING COMPONENTS, attendance and participation will be 10% in your total grade. While we encourage our students to attend the class and or participate in the discussion, the student should not risk his/her health to fulfill this goal. More information about attendance and participation can be found https://sasundergrad.rutgers.edu/degree-requirements/policies/attendance-and-cancellation-of-classes.

Disability Accommodations

In order to receive consideration for reasonable accommodations, a student with a disability must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation." Please see https://ods.rutgers.edu/ or reach out to the instructor for more information.

CIVILITY / COMMUNICATION / CLASSROOM COMMUNITY / SENSITIVE TOPICS

This course is purely based on presentation and discussion. So, **the instructor would like to encourage a respectful communication and supportive classroom community that celebrates diversity.** For the presentation, we want the presenters and participants to be aware of sensitive and uncomfortable topics, language or image.

ACADEMIC INTEGRITY POLICY

Rutgers University takes academic dishonesty very seriously. By enrolling in this course, you assume responsibility for familiarizing yourself with the Academic Integrity Policy and the possible penalties (including suspension and expulsion) for violating the policy. As per the policy, all suspected violations will be reported to the Office of Student Conduct. Academic dishonesty includes (but is not limited to):

- Cheating
- Plagiarism
- Aiding others in committing a violation or allowing others to use your work
- Failure to cite sources correctly
- Fabrication
- Using another person’s ideas or words without attribution, including re-using a previous assignment Unauthorized collaboration
- Sabotaging another student’s work

If you are ever in doubt, consult your instructor.

If you have any questions, please visit the Rutgers University website on Academic Integrity: http://nbacademicintegrity.rutgers.edu/

STUDENT SUPPORT AND MENTAL WELLNESS

Rutgers University provides the following resources to support students in their academic success and mental wellness.

- Student Success Essentials: https://success.rutgers.edu
- Student Support Services: https://www.rutgers.edu/academics/student-support
- The Learning Centers: https://rlc.rutgers.edu/
- Rutgers Libraries: https://www.libraries.rutgers.edu/
- Bias Incident Reporting: https://studentaffairs.rutgers.edu/bias-incident-reporting
- Office of Veteran and Military Programs and Services: https://veterans.rutgers.edu
- Student Health Services: http://health.rutgers.edu/
- Counseling, Alcohol and Other Drug Assistance Program & Psychiatric Services (CAPS): http://health.rutgers.edu/medical-counseling-services/counseling/
- Office for Violence Prevention and Victim Assistance: www.vpva.rutgers.edu/