USC Viterbi School of Engineering
Ming Hsieh Department of Electrical Engineering

Course Announcement

EE/CSCI 451: PARALLEL AND DISTRIBUTED COMPUTATION

TTH 200-320, LAB/DISCUSSION F 200-320

FALL 2024

INSTRUCTOR: VIKTOR K. PRASANNA

The course will focus on broad principles of parallel and distributed computation with applications in Data Science and Machine Learning acceleration. The Lab associated with the course will illustrate the principles through parallel programming examples.

Prerequisite: (EE 355x or CSCI 201L) or consent of the instructor.
Text: Introduction to Parallel Computing, Second edition, Grama, Karypis, Kumar, Gupta, Addison-Wesley.
Course Grade: based on home works, class participation, parallel programming assignments, midterm(s), final and project.

Course Outline:

1. **Architectural Principles for Application Developers:** 1. Pipelined processor (CPU) organization: data and control hazards, ILP, out of order execution, multithreading. 2. Memory systems: cache organization, impact on software performance, locality, multithreading. 3. Communication: static and dynamic networks, communication costs, packet routing techniques. 4. Concurrency and Coordination: comm. costs 5. GPU architecture and execution model.
2. **Parallel Programming Models:** 1. Shared Address Space Programming: shared variables, coordination, Pthreads, OpenMP. 2. Message Passing Programming Model: send receive primitives, blocking and non-blocking commands, collective operations, Cannon's algorithm. 3. Data Parallel Programming Abstraction of GPUs: SIMT execution model, CUDA programming model, Illustrative examples and application mapping, optimizations, OpenCL.
3. **Parallel Algorithm Design and Analysis:** 1. Design Techniques: tasks and dependencies, decomposition techniques, parallelization strategies, data distribution. 2. Analytical Models for Parallel Systems: LogP, BSP, PRAM. 3. 4. Limits on achievable performance, Amdhal's Law, Gustafson's Law, weak and strong scalability, work optimality.
4. **Cloud, Big Data:** 1. Cloud as a computing platform: Large data sets and organization, computational characteristics. 2. Cloud programming models: frameworks, illustrative examples. 3. Map Reduce parallel programming model: Hadoop, example parallelization. 4. Software performance benchmarks: peak performance, sustained performance, LinPack, bandwidth benchmarks.
5. **Accelerating Machine Learning and Data Science Kernels and Applications:** 1. Data Science Kernels: Parallel search and sorting, throughput optimization, multi-dimensional search, decision trees. 2. Communication Primitives for Data Science: broadcast and all to all, communication costs on various topologies. 3. Parallelizing deep learning models: training and inference, data parallelism and model parallelism. 4. Parallelizing graph neural networks, sampling techniques. 5. Benchmarks: Top500, Green500, Graph500, MLPerf.
6. **Systems for Machine Learning:** 1. DNN computations: data and model parallelism, computational requirements, 2. Novel ML applications and their computational aspects: GNN, LLM, GAN, and tensor decomposition. 3. Parallel algorithm design for ML systems: convolution as matrix multiplication, communication and coordination requirements. 4. Accelerating Reinforcement Learning (RL): accelerated primitives, parameterization, illustrative examples. 5. Example systems: PyG, DGN, mapping, performance analysis.
7. **Current Topics:** Heterogeneous Computing, Accelerators, Spatial Computing (FPGAs), advances in parallel programming models, Examples, oneAPI, Sycl, etc.

Professor Viktor K. Prasanna
Email: prasanna@usc.edu
Ext: 0-4483
Office: EEB-200C

USC Viterbi School of Engineering
Ming Hsieh Department of Electrical Engineering

Course Announcement

EE 451
Some student Course Projects completed in earlier semesters

Parallelizing Hessenberg reduction on real square matrices
Multi-core Accelerated AlphaZero using Adaptive Parallelism
Evaluation of Parallel Gradient Descent Methods
Triangle counting on large graphs using SpMM on GPU
Accelerating CNN training using the Log Number System (LNS)
Ray tracer with Pthreads and CUDA: Evaluating GPU performance on control-intensive applications
Evaluating the performance of Cerebras ML accelerator
Spatial Separable Convolutional Neural Networks Parallelization and Acceleration
Parallel first-order logic inference
CUDA Acceleration and Memory Optimization for Transformer Attention
Accelerating the Application of Gabor Filter Banks to Images using GPU
Parallel Genetic Algorithm to solve Traveling Salesman Problem using MapReduce
Accelerated Matrix Factorization using CUDA
Parallelization of Fast Fourier Transform
Parallel Implementation of CNNs for object detection
Evaluation of LLM inference on parallel platforms
Implementation and Analysis of Parallel Delaunay Triangulation
Comparing HW and SW Acceleration for Batch Gradient Decent Algorithms
Parallelizing CNNs on AMD Neural Processing Units (NPUs)
Implementation and analysis of Graph Neural Networks on Intel Meteor Lake Platforms
Implementation and Analysis of Parallel Algorithms for the Maximum Flow Problem in a Network

Professor Viktor K. Prasanna
Email: prasanna@usc.edu
Ext: 0-4483
Office: EEB-200C