Course Description

There is a tremendous need for high performance in the execution of many Machine Learning (ML) algorithms. Until recently, increasing degrees of parallelism in hardware offered a steady source of increased performance. However, with the end of Moore's Law scaling, significant further performance improvements from increased hardware parallelism are not expected. Instead, customization of hardware and other means to achieve higher efficiency with the same number of transistors on chip will be required. Algorithm-architecture co-design will become increasingly important, including automated compiler techniques for effective mapping of algorithms to architectures. This seminar course will cover a range of research issues pertinent to high-performance machine learning in the post-Moore era.

The papers to be studied in this seminar cover three topics:

A) Analysis/optimization of ML algorithms

B) Frameworks/compilers for high-performance ML

C) Architectures for ML

We expect that participation in this seminar by students and faculty with diverse research interests spanning topics in parallel algorithms/compilers/architecture topics with a focus on Machine Learning will be conducive to identifying and discussing cross-cutting research problems worthy of further study by Masters and Ph.D. students.

Students may register for 1 credit (paper presentation) or 2 credits (paper presentation and project).