I’m a first year graduate student in computer science. I just transferred from civil and environmental engineering department this January. My focus is machine learning. I have done some research work on data analysis but not too heavily on programming or in the more “pure CS” sense. I took a class in civil engineering department on information systems that brushed lightly on parallel computing. Knowing that in machine learning, especially for large scale problems, it is important to be able to perform parallel computing. So I really wish to learn about parallel computing and be able to apply it to my research.

When running large scale optimization and simulation problems in Transportation (where I used to study in), one run usually take a long time. A friend of mine used some basic level of parallel computing to speed up the process and received the results faster than running on one computer.

The problem he needed to solve was to simulate several traffic control schemes over a simulated city and select the case with best performance. These are very complicated models and take a very long time to run. Running it using simply method would take several hours, and it is very vulnerable to errors and bugs, as they will terminate the simulation and the progress will be invalid. After using parallel computing, he was able to speed the whole process by around 40% with another computer, and achieve essentially same results. From my conversation with him, he used python, and pp and dispy libraries. He only used another machine that didn’t have very strong hardwares, so the boost in computational power wasn’t great. I believe if he had proper machines and set up the whole problem more carefully he will have a great boost in performance. In the future, if larger scale problems are encountered, he can use similar method to solve them.