CS 267 - ASSIGNMENT 0

Bio:
I am currently a first year graduating student, majoring in Systems Engineering. My research interests lie in machine learning and Big Data for social scientific experiment. Through this course, I’d like to learn more about parallelism at the software level, the algorithms, and programing model in the distributed and cloud computing environment.

An application problem for which parallel computing has been used:
Processing big data – Real-time GPS locations and users’ feedbacks from mobile devices.

Last semester, we ran a big experiment [1] studying about the public transportation, MUNI in San Francisco. One of the objectives of this experiment was to implement a survey platform on the users’ handheld devices. There were about 800 participants and we collect GPS locations of each user every 30 seconds as well as the user’s feedbacks once or twice a day.

At the sign-up phase, the server was down due to high traffic demand. So we turned on the load-balancing feature on Amazon AWS. As for the parallel platform, we tried to use the distributed computing model, which did most of the computations on the mobile devices and pushed the results to the servers.

The system architecture consisted of two main components. One was the Android app and the other was the server, which was written in Python and mostly consisted of APIs to update and request data. The application itself sitting on either the mobile or the server did not “scale.” We had to use the commercial scaling services, Amazon AWS Elastic Cloud, and trusted that the Postgres database server implemented the parallel computing model and had its own scaling mechanisms.
The advantage of using the Elastic Cloud services was that it automatically balanced the loads on our servers and we didn't have to pay much attention to it. However, since we only had one big central database, it introduced the bottleneck and there were nothing we could do about it. Therefore, I would love to learn more about how to improve the system at the software level for a next bigger experiment.

Now we are at the analyzing data phase and I’d like to use Hadoop to process our collected data and do some machine learning. It’d be nice if I can apply what I’ve learned in this course.