This project extends work in a workshop paper by Li, et. al that was submitted to HotOS '15 (the paper has not yet been published officially).

Motivation:

In parallel computing, resource allocation is necessary for executing large-scale processing jobs. A job is divided into tasks which are then run in parallel. The traditional scheme has been for the scheduler to assign tasks according to resource availability. This approach is problematic for two reasons: the resource requirements for a task are inherently uncertain, and the amount of available resource visible to the scheduler is also uncertain. In order to address these issues, a task allocation approach, instead assigning tasks to resources, was suggested by Li, et. al. This strategy is similar to how TCP probes and fills up available bandwidth on a bottleneck link. Such an algorithm has not been previously implemented before and hence requires
significant experimentation and testing in order to determine proper thresholds and parameters and task allocation logic.

My project:

Preliminary task allocation algorithm pseudocode is rough and given in the workshop paper.

In the beginning, I will examine Hadoop source code, existing scheduling algorithms, read up on TCP, and learn performance profiling tools in order to familiarize myself with the background necessary for this project.

When machines for a Hadoop cluster arrive at the end of the third week, I will begin experiments in order to make initial evaluations. Such experiments will require thinking about different test cases to optimize, as listed in the workshop paper, and performance profiling while running multiple tasks on Hadoop.

The ultimate goal of my project will be to contribute code to an implementation of the task allocation algorithm.

Timeline and Deliverables:
Week 1: Research background, Hadoop and source code

Week 2: Research background, TCP and existing scheduling algorithms

Week 3: Research performance profiling tools

Week 4: Design initial evaluation experiments and try running them on Hadoop cluster

Week 5: Compare experiment results with existing results

Week 6: Try writing initial code for scheduling algorithm, along with graduate students

Weeks 7-12: See what happens with experiments and change code accordingly; a more clear list of tasks will be constructed and split with graduate students

End of semester: Final report