# Compiling Neural Networks with Dynamic Resource Adaptation: **Project Milestone Report**

Patrick Coppock
pcoppock@andrew.cmu.edu

Yuttapichai (Guide) Kerdcharoen
ykerdcha@andrew.cmu.edu

https://pnxguide.github.io/15745_project/

## Major Changes

We have made no major changes to our project goals or methods.

## Progress

We have implemented FlexGen policy search (i.e., cost model and policy search via linear programming) and can run it once at the beginning of inference jobs. We are now trying to evaluate FlexGen results based on the policy search we implemented. This implies that we have accomplished halfway through our 75% goal.

Toward the 100% goal, we are now implementing dynamic block scheduling by dropping batches as memory becomes scarce. We are also generating the synthetic environment to evaluate our dynamic block scheduling.

We did not quite complete our original milestone goal (which was optimistic) but are on track to wrap it up soon.

## Surprises

Our first task, implementing the cost model and policy search, was more complex than we anticipated.

While not a surprise, our evaluation plan is a concern. Because we are extending an existing scheduling system to be more resource-agile, evaluation is less straightforward than it would be if we were writing a new compiler pass. We need to come up with a specific evaluation plan.

## Revised Schedule

We are a week or so behind our schedule. We are on track to completing the conservative goal. Additionally, we believe we can still complete one of our stretch goals. At this point, we will plan to complete only stretch goal B, which is to run full policy search throughout block inference, tweaking policy as allowed. Implementing this stretch goal may allow our system to utilize the hardware more efficiently than our conservative goal does.

## Resources Needed

The one resource we listed in our proposal, a large GPU machine, we have been successfully using. While its availability is somewhat spotty, our progress has not been impacted.