# CS-E4660 - Advanced Topics in Software Systems

Planning Document

Linus Jern

October 30, 2024

## Problem Definition

The problem for this project is the dynamic management of models and ML service provisioning with a focus on system latency given the accuracy constraints of any given request. This problem domain is quite complex, so some clarification is needed. In this context, *dynamic management of models and ML service provisioning* means the dynamic control of:

- Model size
- Resource Utilization

An example of a system where this might be worth exploring is a system with strict latency requirements, but quite loose accuracy requirements. In such a system, instead of letting each inference node sequentially handle the incoming queue of requests, the node might choose to preempt the queue and take the next job if the system resource utilization of it's own resources are low enough. In addition to preempting the sequentiality, the node might choose a smaller, and thus faster with less resources, model if the circumstances allow.

What are the circumstances that would allow this compromise? If the task is of a type that has a accuracy requirement or threshold which is low enough to compromise for a smaller and/or less accurate model, this preempting can occur.

See figure Figure 1, for a rough drawing of the idea space.

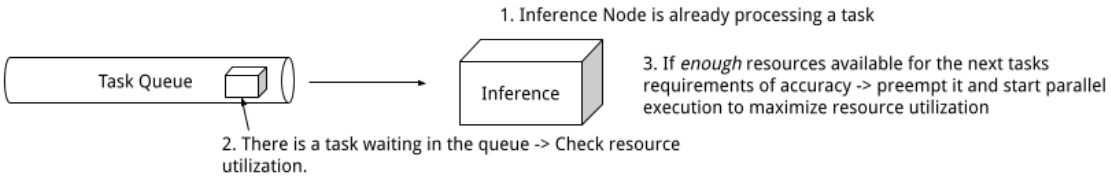


Figure 1: Rough Drawing of the Situation