InferLine: Prediction Pipeline Provisioning and Management for Tight Latency Objective

Serving prediction pipelines spanning multiple models and hardware accelerators is a key challenges in production machine learning. **Optimally configuring these pipelines to meet tight end-to-end latency goals is complicated by the interaction between model batch size**, **hardware accelerator choice** and **variation in query arrival process**.

A system to provision and execute ML prediction pipelines subject to end-to-end latency constraints by proactively