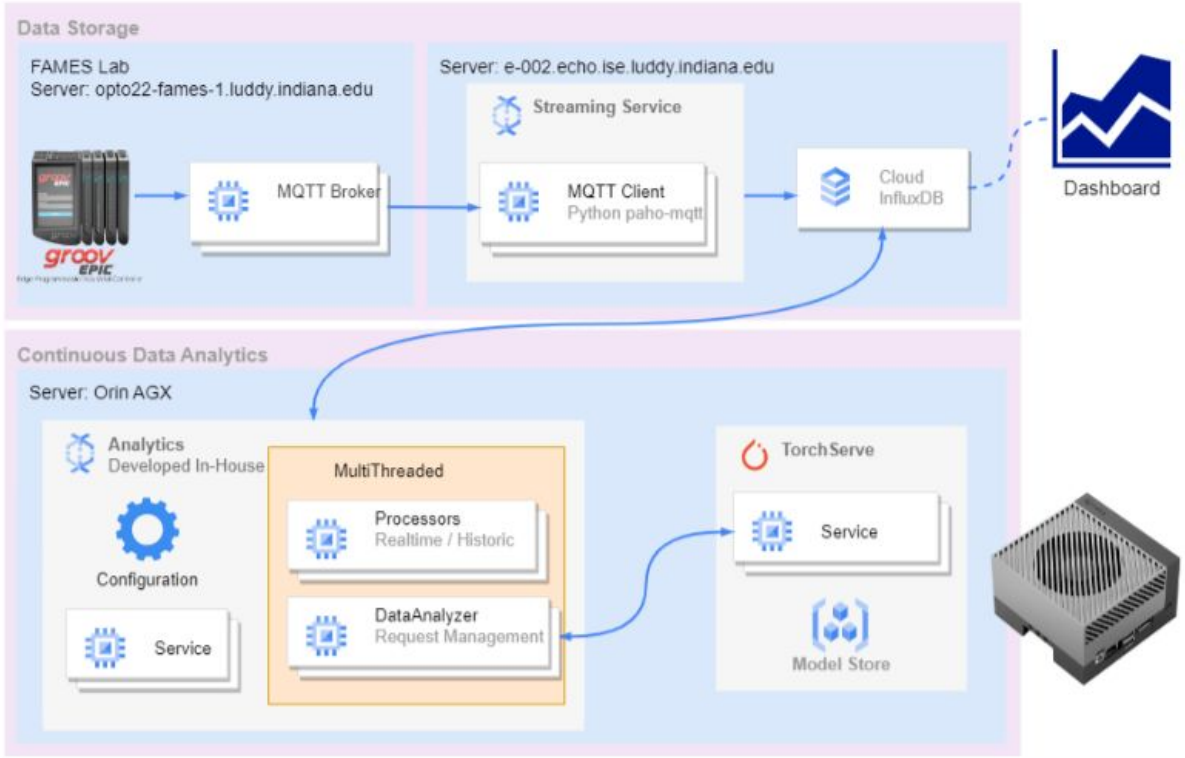
E517: Project - Quantifying Performance and Energy Efficiency of a Real Time Data Analytics System

Abdul Rehman

Real Time Data Analytics System



Problem Statement

We want to optimize.

Problem Statement

We want to optimize.

Before we can optimize we must measure.

Problem Statement

We want to optimize.

Before we can optimize we must measure.

To optimize for	We must measure
Energy	Power
Performance	Inference Time

Problem Statement

We want to optimize.

Before we can optimize we must measure.

To optimize for

Energy

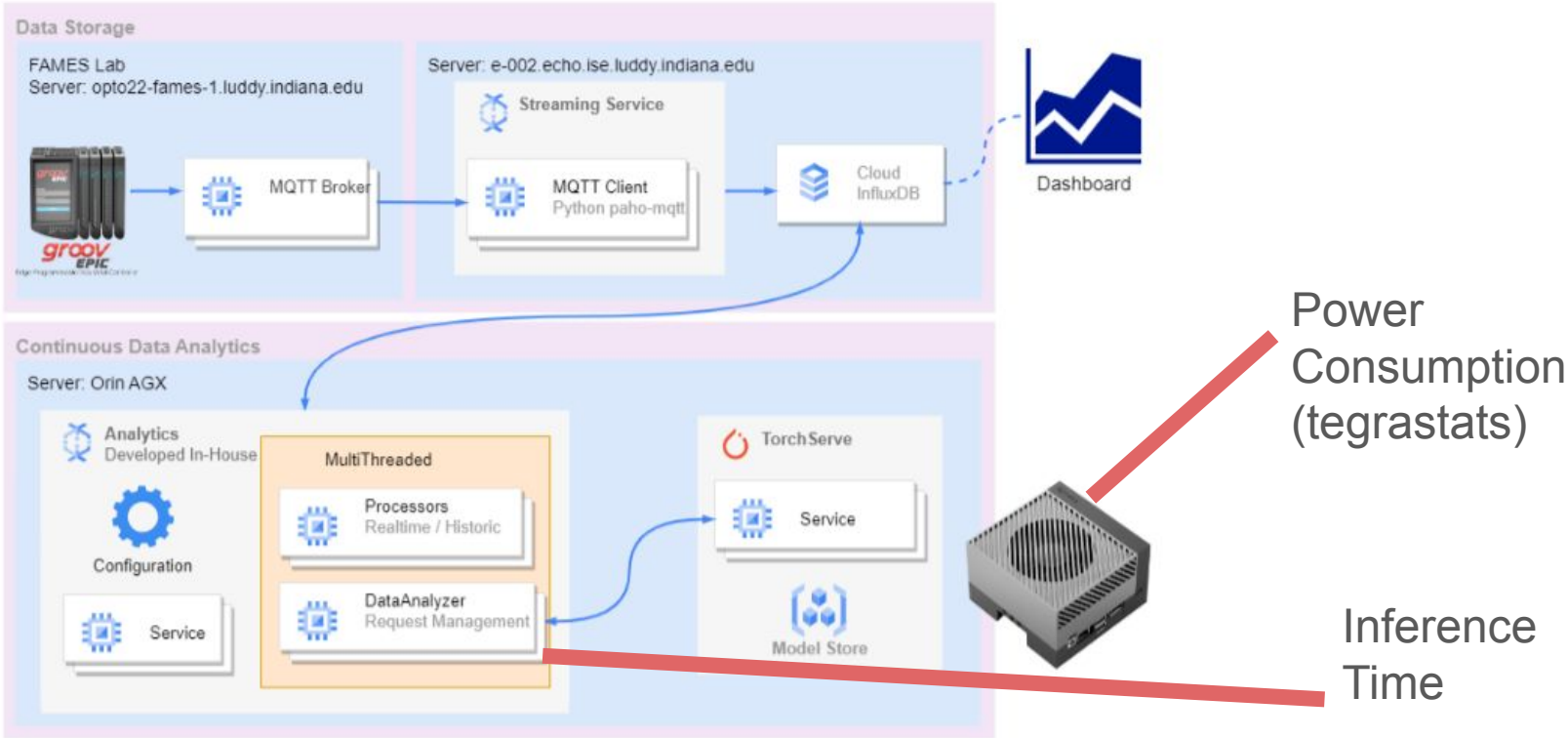
Performance

We must measure

Power

Inference Time

Measurement



Results: Inference Time

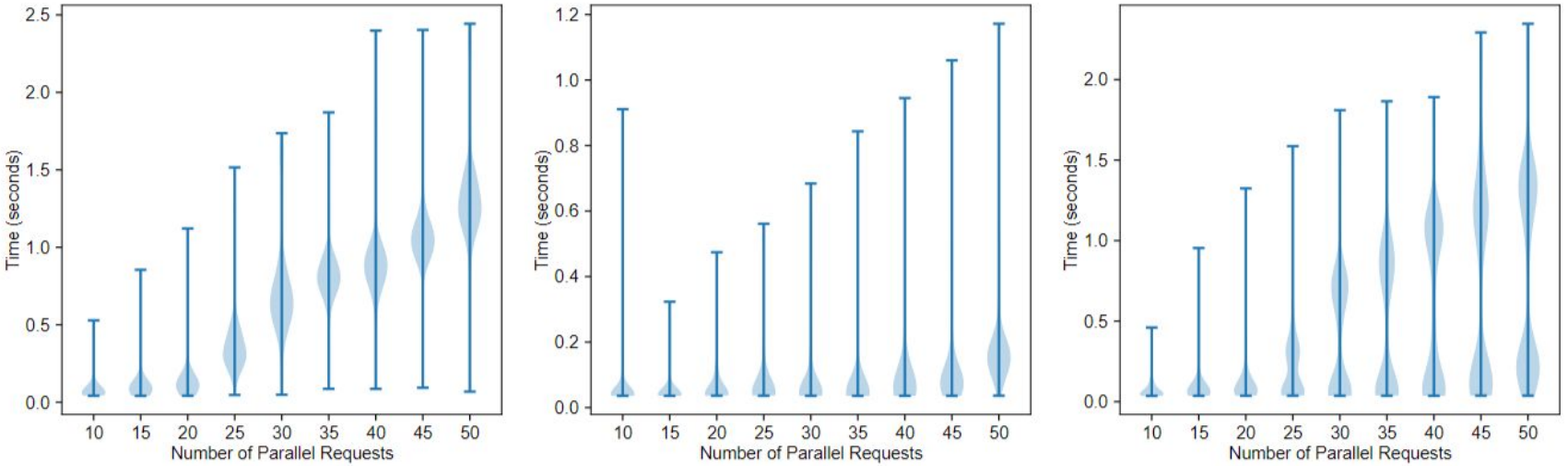
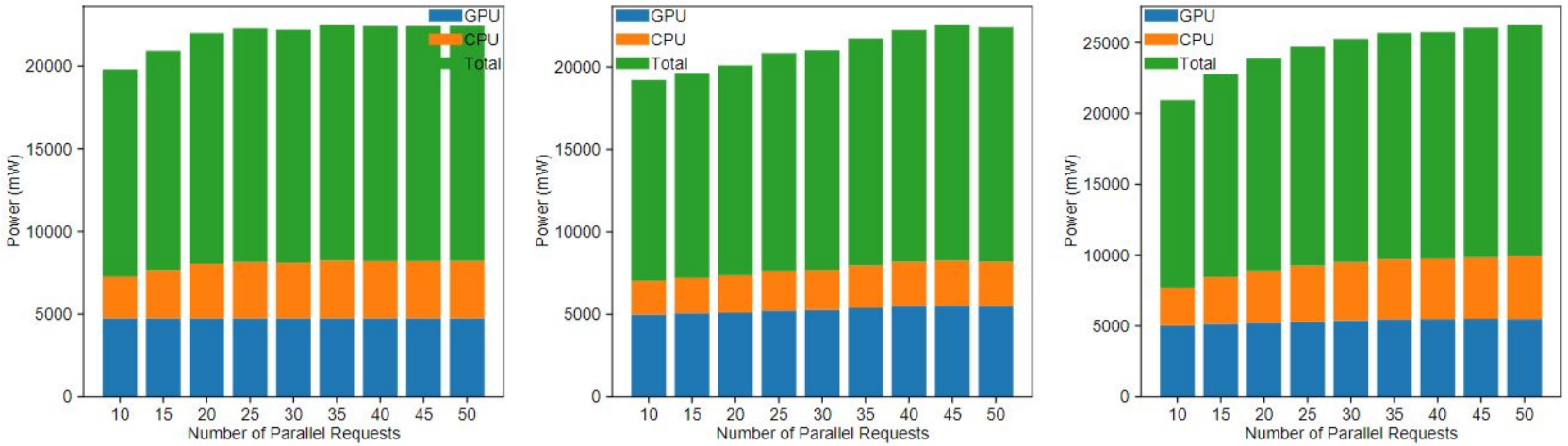


Figure 4: Inference Times

Results: Power Consumption



(a) ARIMA (b) ARNN (c) BOTH

Figure 3: Power Consumption

Key Findings

• System is stable for GPU centric models as opposed to CPU centric models in terms of tail latency.

• GPU is power efficient as opposed to CPU.

• Point of inflexion is reached rapidly for CPU centric models as opposed to GPU centric models.

Discussion

Possible ways to optimize

- Selective throttling based on the inflexion point for CPU and GPU specific requests.
- Build GPU centric models for existing CPU centric models.

Thank You!