CS 744: INFAAS

Shivaram Venkataraman

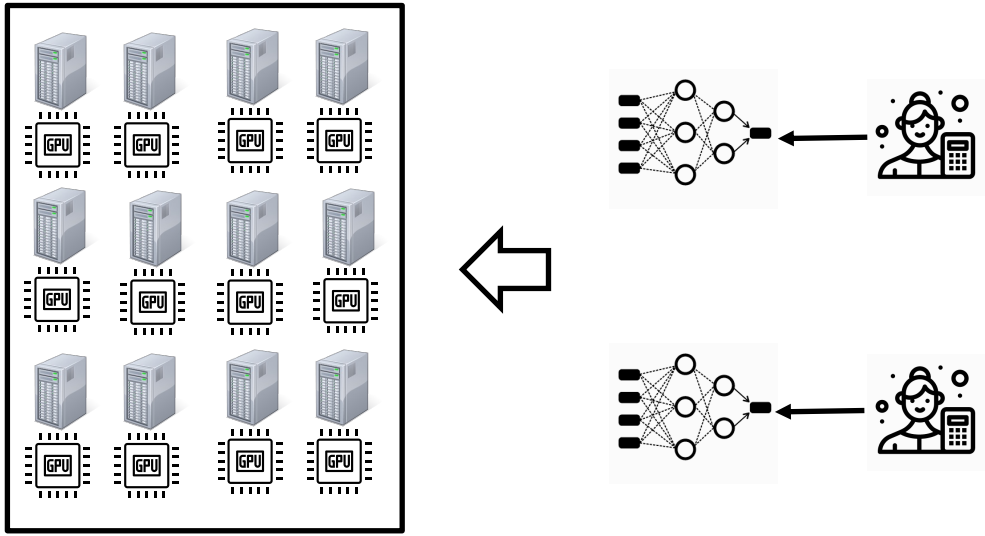
Spring 2024

ADMINISTRIVIA

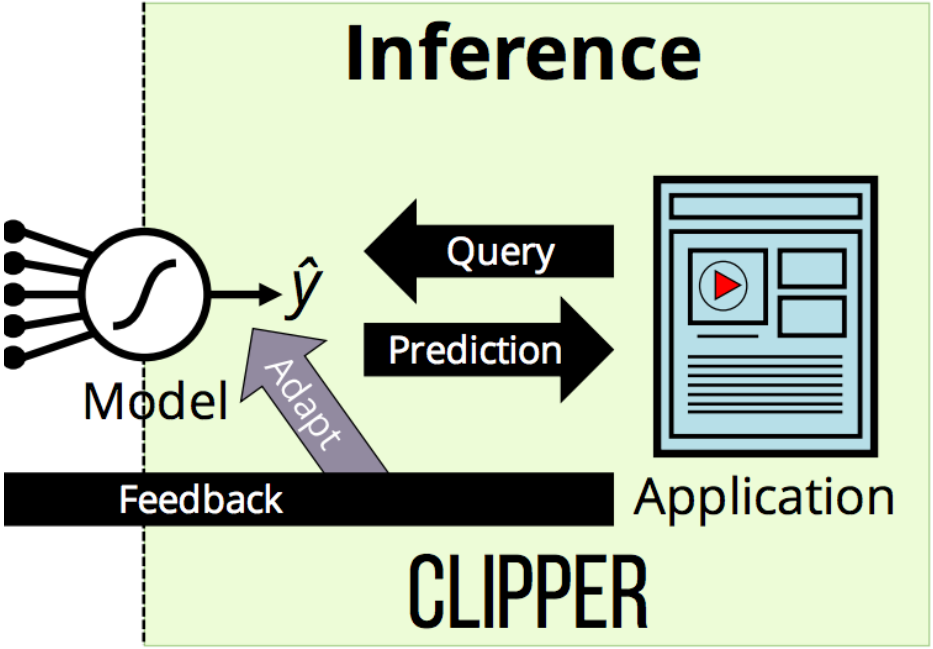
- Course project
  - Introductions due March 8th

- Midterm Exam
  - In class on March 14th
  - Includes everything from beginning to the end of scheduling (including INFaaS)
  - Sample papers on Piazza

MACHINE LEARNING: TRAINING, INFERENCE



ML INFERENCE: MODEL-FUL?



MOTIVATION: MODEL VARIANTS!

Diverse Applications

Application	Accuracy	Latency	Cost
Social Media	High	Medium	Low
Visual Guidance	High	Low	High
Intruder Detection	Low	Low	Low

Heterogeneous environments

Diverse model variants

Variant (hardware, framework)	Lat. (ms)	Req/s	Cost ($/s)
A (4 CPUs, TensorFlow)	200	5	1
B (1 Inferentia core, Neuron)	20	100	3
C (1 V100 GPU, TensorRT)	15	800	16

QPS	SLO (ms)	#Var. A	#Var. B	#Var. C	Cost ($/s)
10	300	2	0	0	
10	50	0	1	0	
1000	300	0	2	1	

INFAAS: MODEL-LESS SERVING

Insight: Opportunity to pick right model for a query

User specifies performance, cost, or accuracy requirements

System manages model variants at runtime

INFAAS: OUTLINE

Model Variants

API Design & Architecture

Selecting, Scaling Model variants

MODEL VARIANT

Architecture

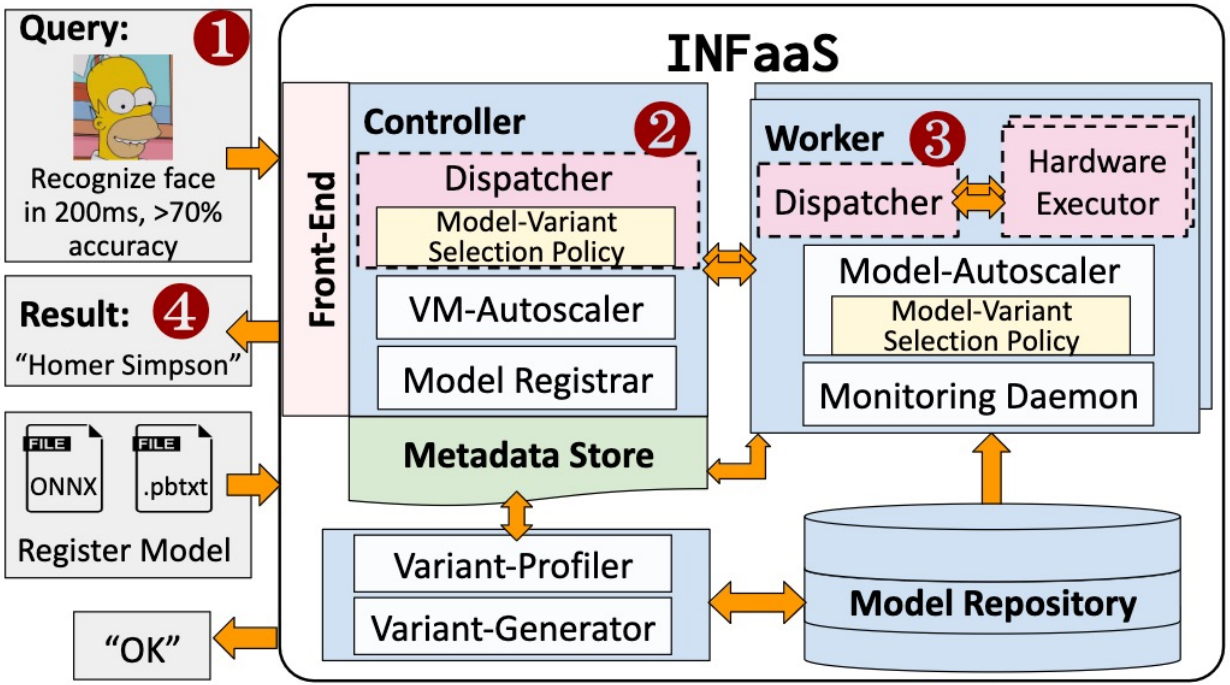
Programming Framework

Graph Optimizer

Batch size

Hardware

```
register_model("ResNet50",ResNet50.pt,valSet,detectFaceApp)
register_model("MobileNet",MobileNet.pt,valSet,detectFaceApp)
```

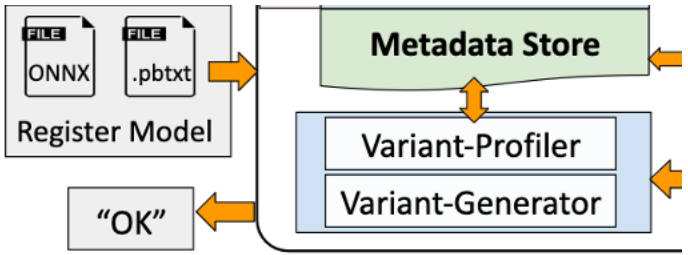


VARIANT GENERATOR

Generates model variants

Accuracy using val set

Profiler used to measure latency, memory use, loading time etc.



MODEL VARIANT SELECTION POLICY

Case 1: Query arrival

Goal: minimize time to select model variant

Approach

- Select least loaded active variant (no loading time)
- If not, select a variant with low loading+inference time and worker with low utilization

MODEL VARIANT SELECTION POLICY

Case 2: Query load change

Options

- Horizontal scaling, replication
- Vertical scaling!

$$\text{Cost}(\delta_{ij}) = C_{ij}(\delta_{ij} + \lambda T_{ij}^{\text{load}} \max(\delta_{ij}, 0))$$

GREEDY HEURISTIC

Prune the search space for variant selection!

- Calculate headroom available on each worker
- How to scale?
  - Estimate cost of horizontal, vertical scaling
  - Compute cost for both proposals
  - Check if it fits in the worker

SUMMARY

DL inference workloads properties

Model-variants search space large

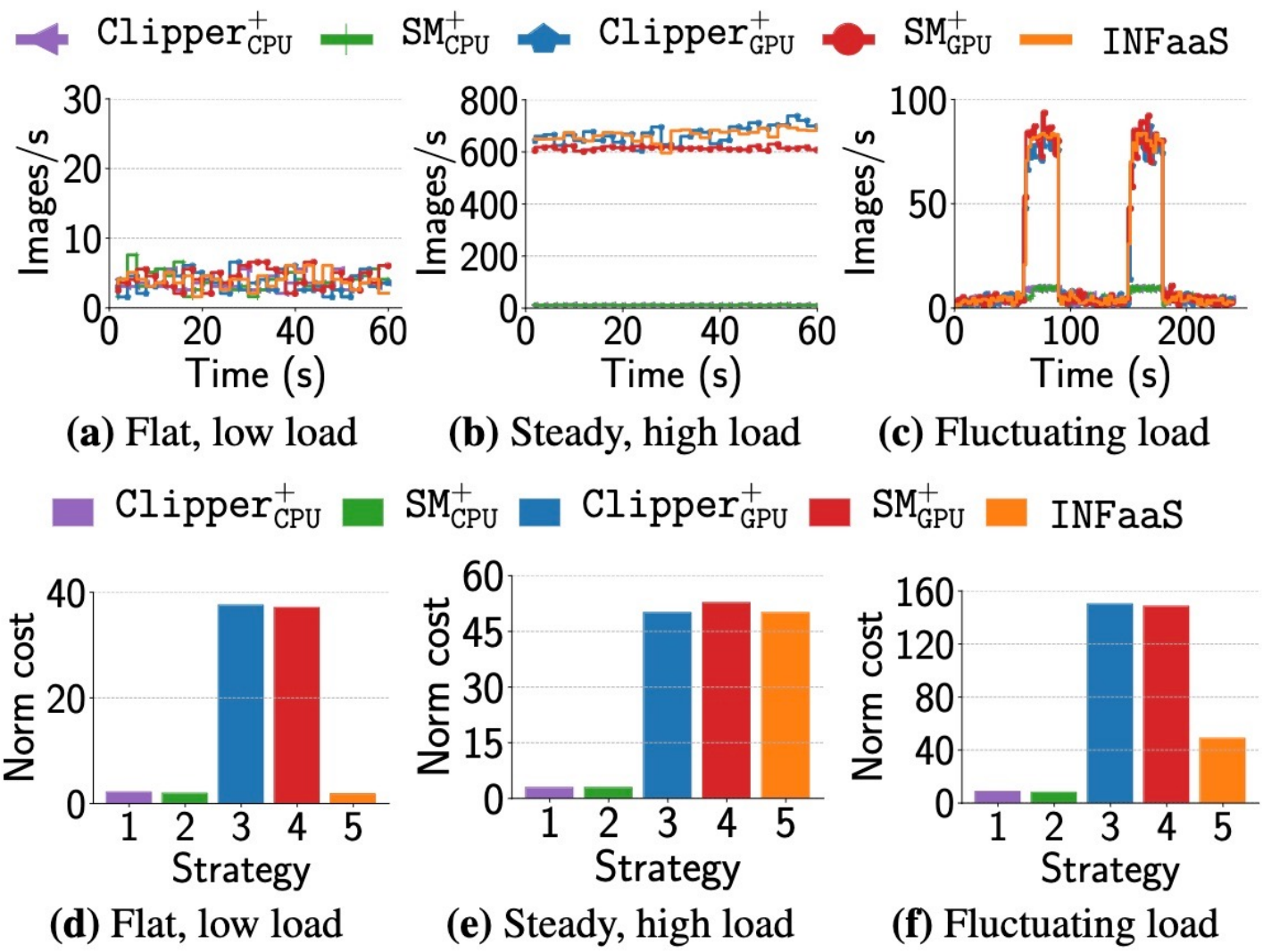
INFaaS: Model-less serving

Scale model variants horizontally and vertically

DISCUSSION

https://forms.gle/gkNTqdrSpLVpa9rg8

List one similarities and one difference between training schedulers like Gavel and inference scheduler like INFaaS



(a) Flat, low load (b) Steady, high load (c) Fluctuating load

(d) Flat, low load (e) Steady, high load (f) Fluctuating load

NEXT STEPS

Next Class: SQL

Course Project Introductions!