

CS 744: SINAN

Shivaram Venkataraman

Spring 2025

ADMINISTRIVIA

Grading updates

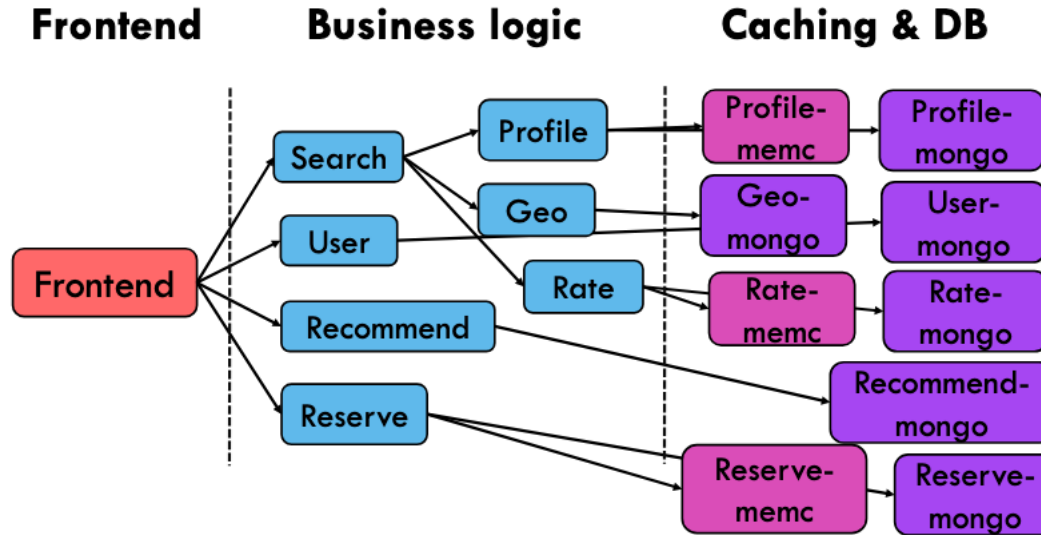
Midterm 2, April 24th

-- Check Piazza

Poster session: May 1st

— More details soon

MICROSERVICES

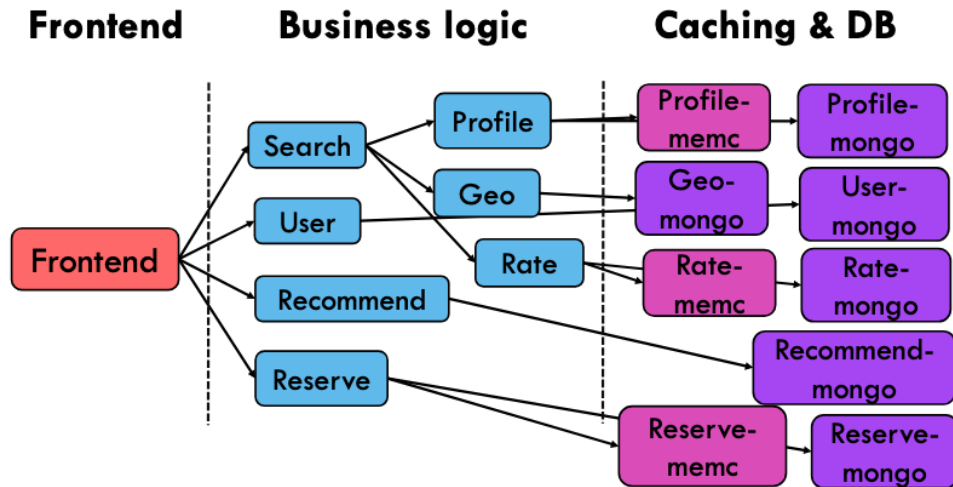


RESOURCE MANAGEMENT CHALLENGES

Dependencies among “tiers”

Delayed queuing

Resource allocation space



SINAN APPROACH

Narrow down the search space of possible allocations.

Given an allocation **use ML** to predict performance

APPLYING ML TO SYSTEMS

Data collection
Agent

CNN Model

Predict tail latency

Boosted Tree

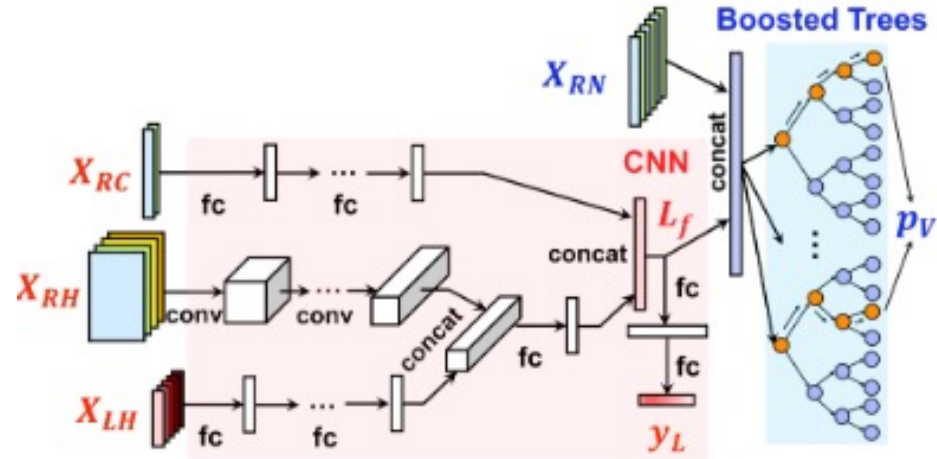
Probability of QoS
violation in future

MODELING CHALLENGES

Goal: Predict latency given a resource allocation

Latency at what time point?

QoS violation: Classification vs. regression

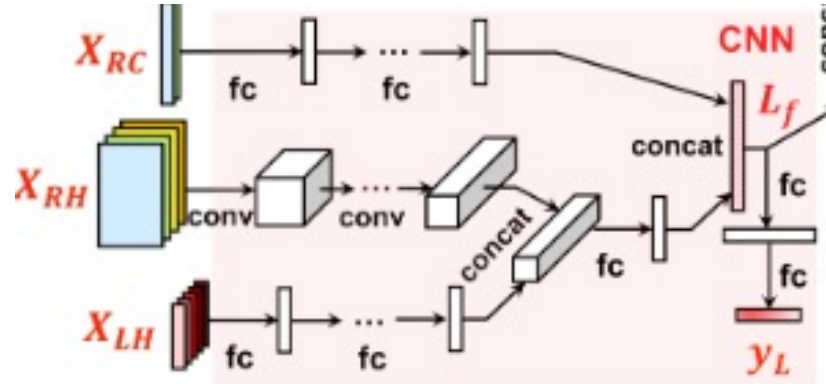


TWO STAGE MODEL: CNNs

3D tensor ("image") resource allocation over time
Dimensions include: microservice tier, resource
(CPU, mem), time

End-to-end latency distribution within time
window (X_{LH})

Resource configuration for next step (X_{RC})



$$\mathcal{L}(X, \hat{y}, W) = \sum_i^n (\hat{y}_i - f_W(x_i))^2$$

TWO STAGE MODEL: VIOLATION PREDICTION

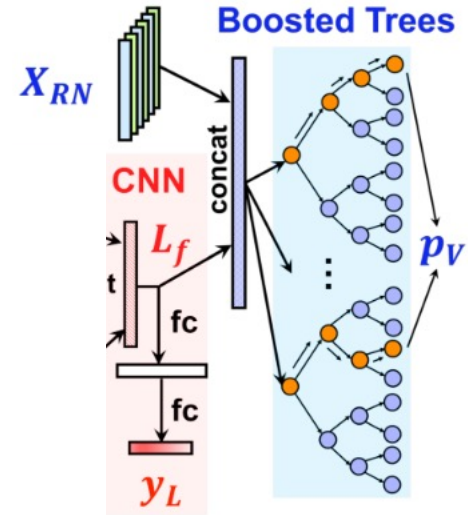
Boosted Trees

Classification problem: QoS violation or not

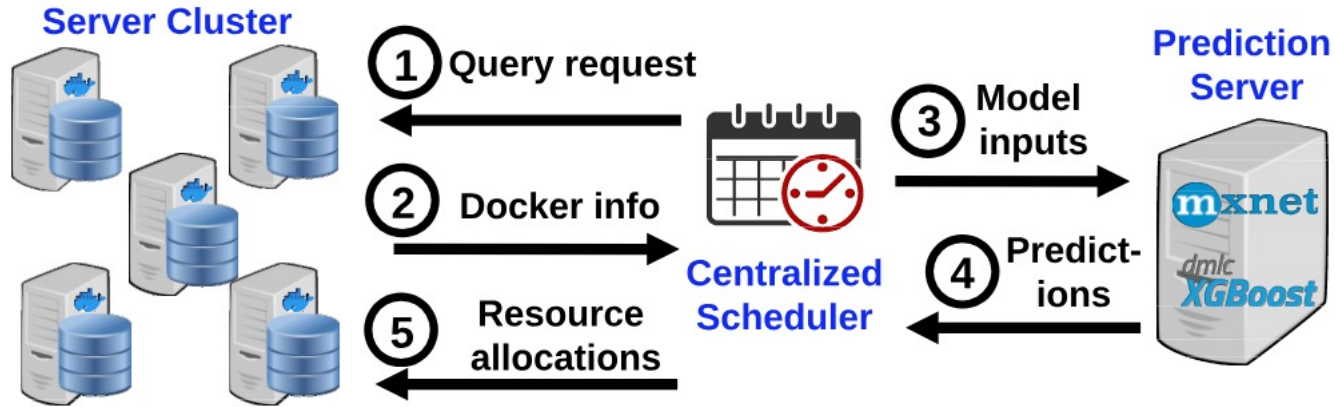
Input: last layer of CNN (smaller in dim than inputs)

Resource allocation

Output: probability of violation



SYSTEM ARCHITECTURE



HOW TO COLLECT TRAINING DATA

Needs to cover a spectrum of application behaviors

Boundary of the resource allocation space. Why?

Multi-arm bandit based approach

- treat each tier as an arm.
- each time select one tier to explore (max information gain)

TRAINING DATA ACTION SPACE

Only allowed a predefined set of operations

- reducing or increasing the CPU allocation by 0.2 up to 1.0 CPU

- increasing or reducing the total CPU allocation by 10% or 30

- upper limit on CPU utilization at each tier

Training data should see some samples which are above QoS

ONLINE SCHEDULER

Exclude operations whose predicted tail latency is high

Use the predicted violation probability to filter out risky operations

Category	Actions
Scale Down	Reduce CPU limit of 1 tier
Scale Down Batch	Reduce CPU limit of k least utilized tiers, ($1 < k \leq N$)
Hold	Keep current resource allocation
Scale Up	Increase CPU limit of 1 tier
Scale Up All	Increase CPU limit of all tiers
Scale Up Victim	Increase CPU limit of recent victim tiers, that are scaled down in previous t cycles

SUMMARY

Applying ML to Systems → new direction

Sinan: Apply ML to microservice scheduling

Open questions

- Model arch – 2 level with CNN, Boosted trees

- Training data – multi-arm bandit based approach

- Model inference – online scheduler with smaller search space



DISCUSSION

<https://forms.gle/xSyIBkPWGedDKYAZ6>

If you add a new service to your graph or make changes to complexity of a service, how will you make sure Sinan handles it ?

How should the scheduler handle cases where the ML model returns incorrect predictions?

NEXT STEPS

Next week schedule

Tue: Luminix (Inference scheduler)

Thu: Midterm 2