MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving

Chengliang Zhang†, Minchen Yu†, Wei Wang†, Feng Yan‡

†Hong Kong University of Science and Technology

‡University of Nevada, Reno

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Machine Learning Model Serving

Deploy a trained model for user requests

- Highly dynamic demand
- Stringent Service Level Objectives on latency

Design objectives

- Serve ML models on public cloud
- Scale to dynamic queries
- Cost-effective
- SLO-aware: e.g. 98% of the requests must be served under 500ms

Challenges & Opportunities

Unique properties of ML serving

- Compute intensive
- Hardware accelerators: GPU, TPU
- Stateless computation

How to reduce over-provisioning?

Cloud services

- Multiple options: IaaS, CaaS, FaaS, MLaaS
- Large configuration space: CPU, memory
- Cost-performance tradeoffs: preemptable, burstable instances

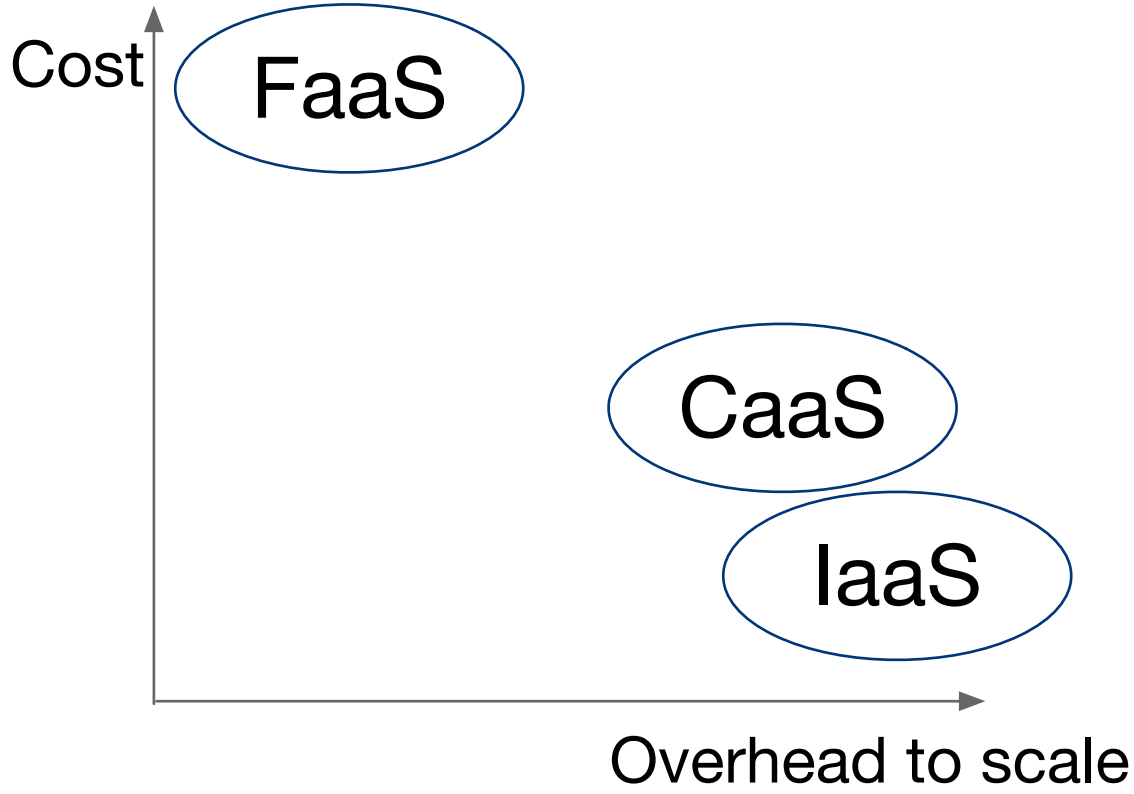
What option to choose?

Cloud Services for Model Serving

Infrastructure as a Service (IaaS)

Container as a Service (CaaS)

Function as a Service (FaaS, serverless comp.)

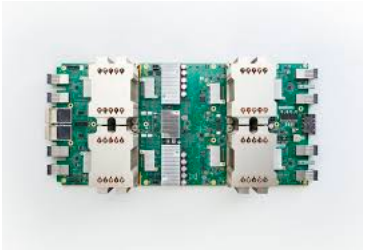


Cloud Services for Model Serving

CPU

GPU

TPU



…

On-demand

Preemptable

Spot instances in AWS
Preemptable VM in Google cloud

Burstable

t2, t3 instances in AWS
f1-micro, g1-small in Google cloud

…

We designed MArk

A scale-to-demand, cost-effective, SLO-aware model serving system on cloud

Compared with AWS's SageMaker, MArk achieves

- Up to 7.8x cost reduction
- Better latency performance

Welcome to our talk!

Day 3, Track II, Machine Learning Applications & System Aspects

- Our insights of IaaS, CaaS, FaaS and their configurations
- Our insights of ML serving on GPUs and TPUs
- How MArk translates our insights into system design
- MArk's provisioning algorithm
- The evaluation of MArk's performance