PennState

Implications of Public Cloud Resource Heterogeneity for Inference Serving

Jashwant Raj Gunasekaran, Cyan Subhra Mishra, Prashanth Thinakaran,
Mahmut Kandemir, Chita Das

Sixth International Workshop on Serverless Computing (WoSC6)

Dec 8, 2020

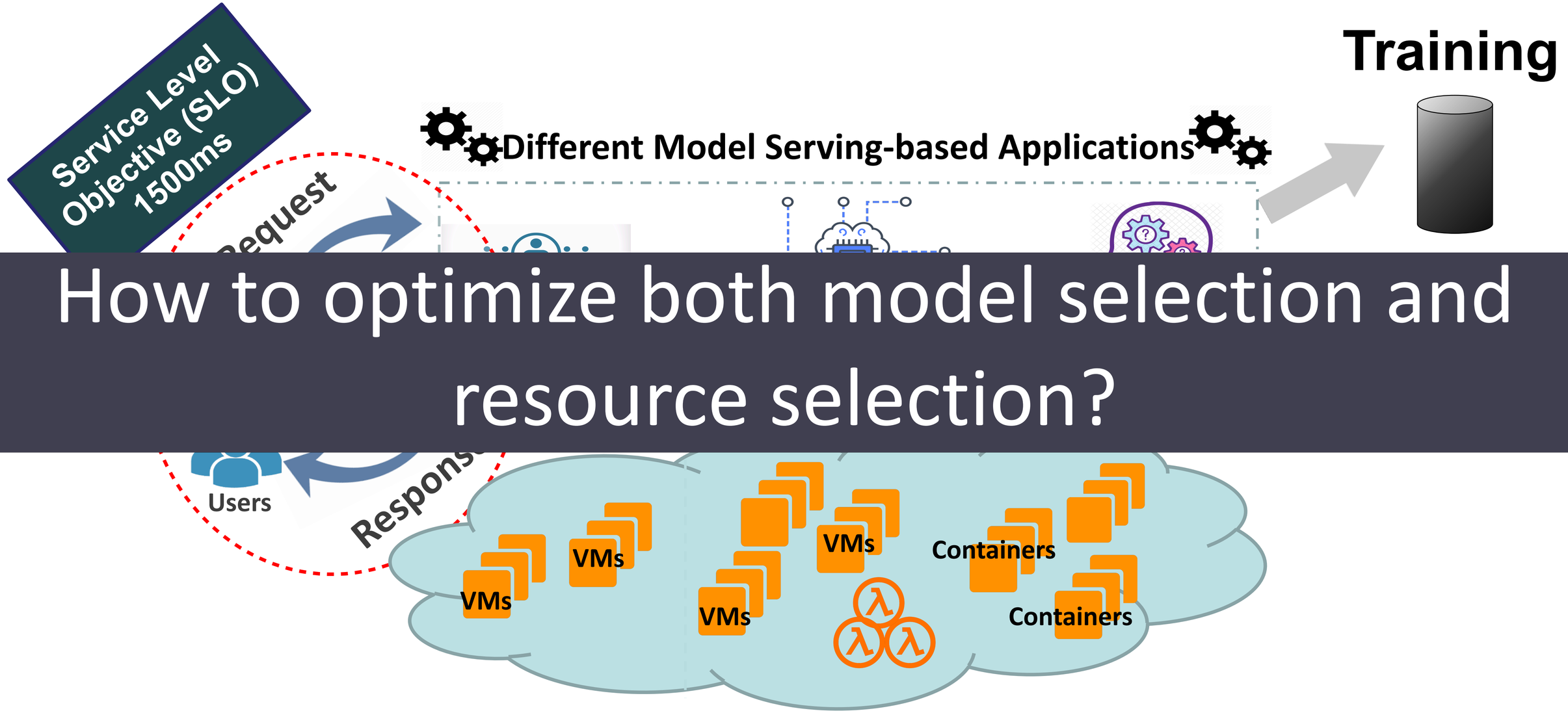
PennState
College of Engineering

PennState
High Performance
Computing Lab

EXECUTIVE SUMMARY

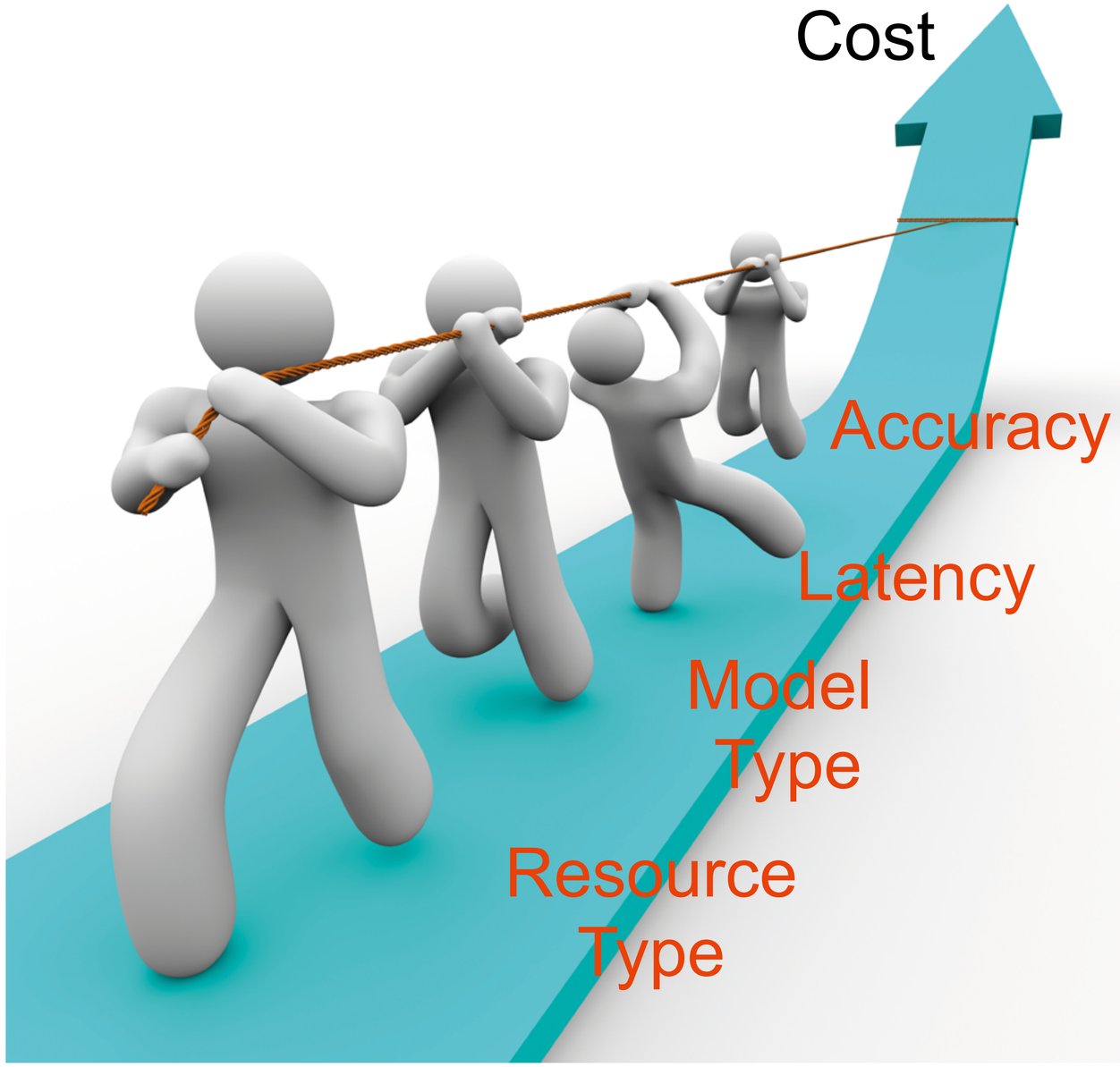
TENANTS		PROVIDERS
Faster Response Times	→	Multiple Service Offerings
SLO violations, Variable Cost	↔	Model Selection Resource Selection
Low Cost, SLO-aware	↔	Self Managed Automated Framework

Model Serving Hosted on Cloud

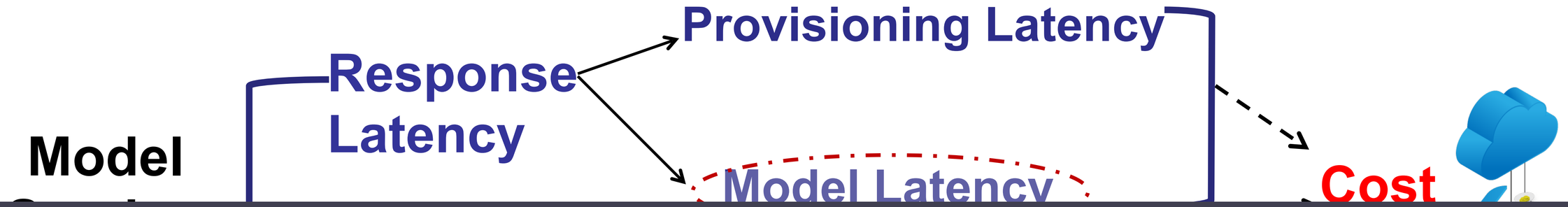


How to optimize both model selection and resource selection?

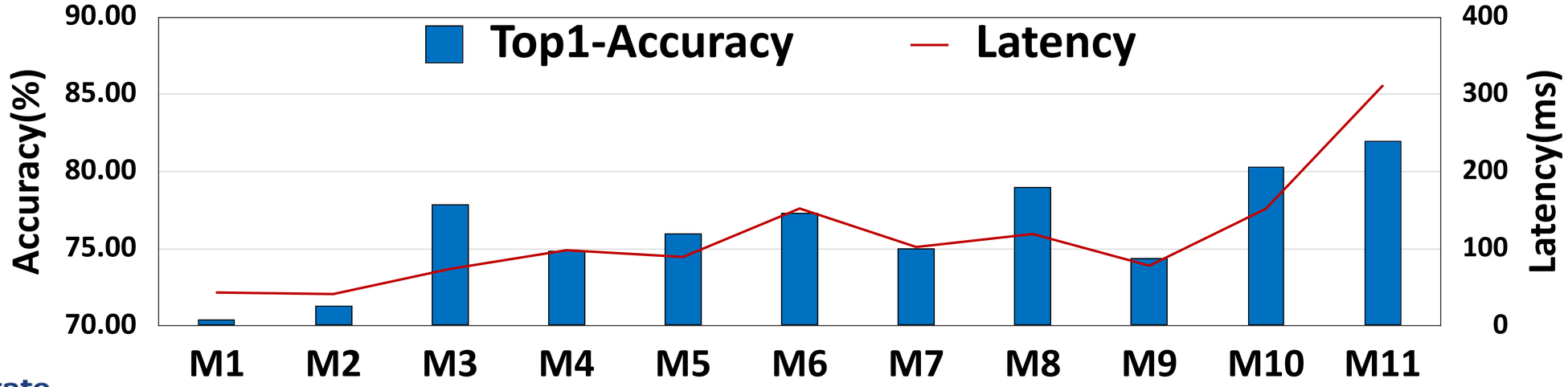
Model Serving in Public Cloud



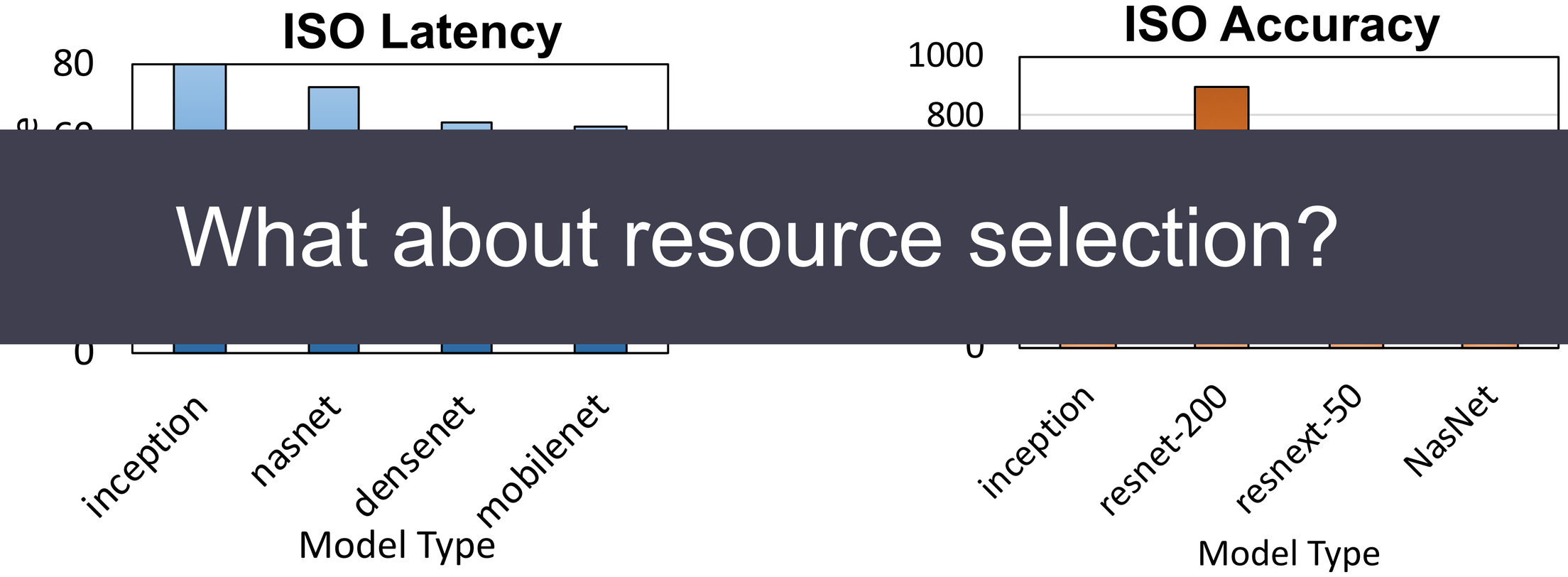
Model Serving Requirements



Model Serving Challenges?

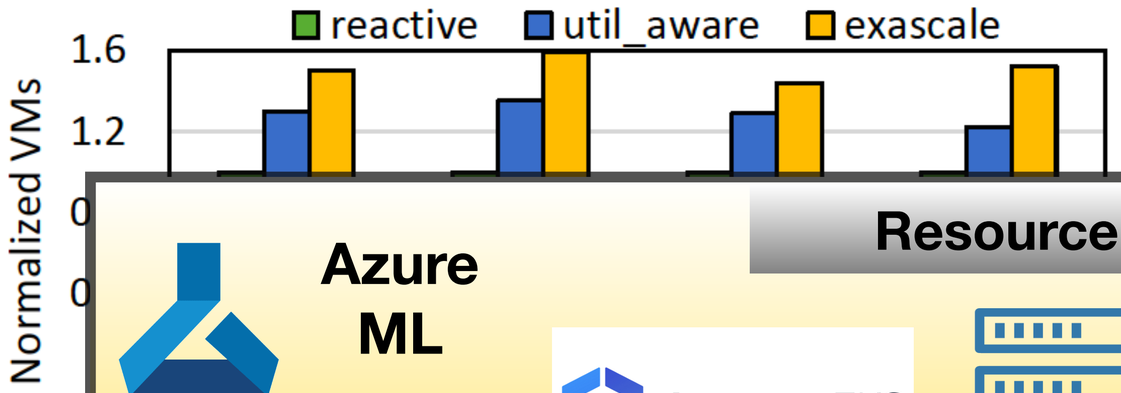


Model Selection



What about resource selection?

Analyzing Prior Works



SLO violations **#VMS**

Resource Offerings from Cloud

Azure ML

Amazon SageMaker

Amazon EKS

Baremetal

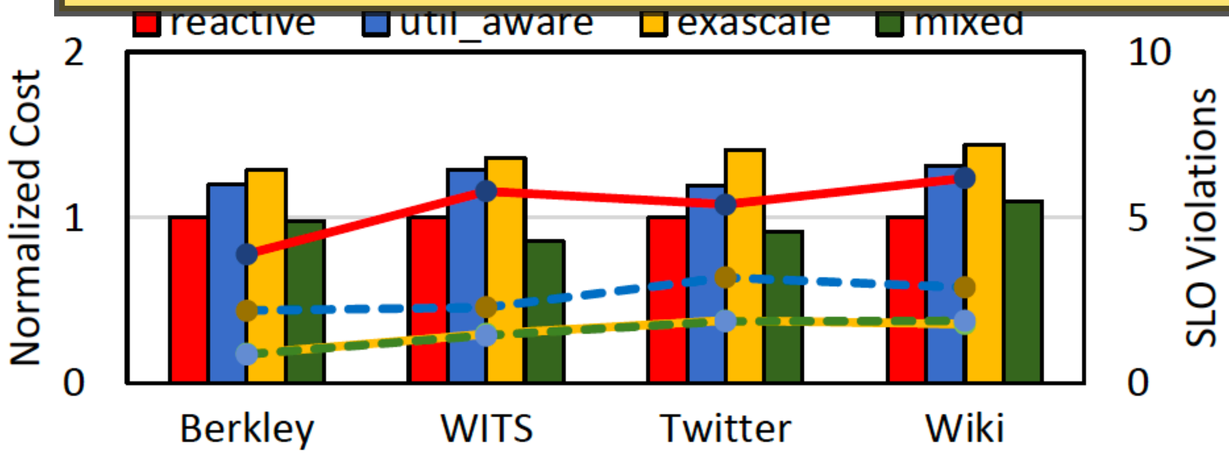
vm

Cloudlets

docker

Serverless Functions

unikernel

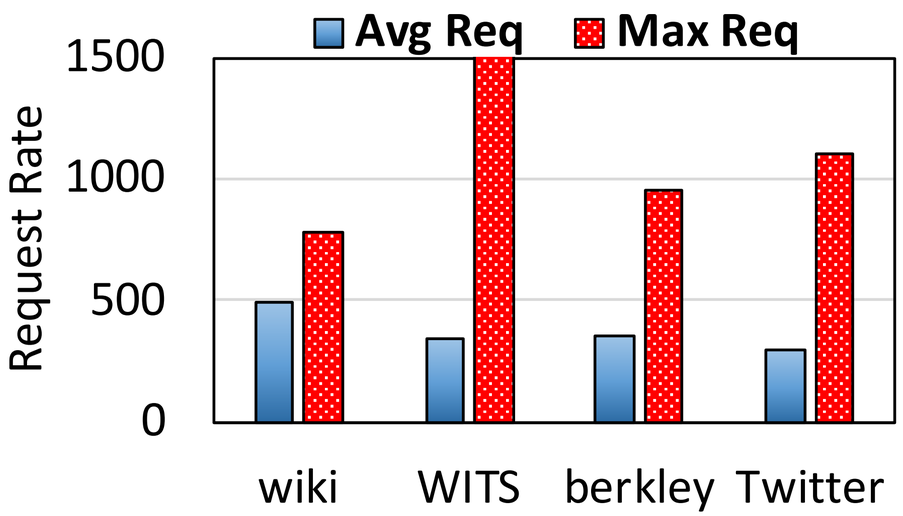


Cost of Different Policies

VM-Cost **#VMS** **Mixed**

Challenges with Serverless

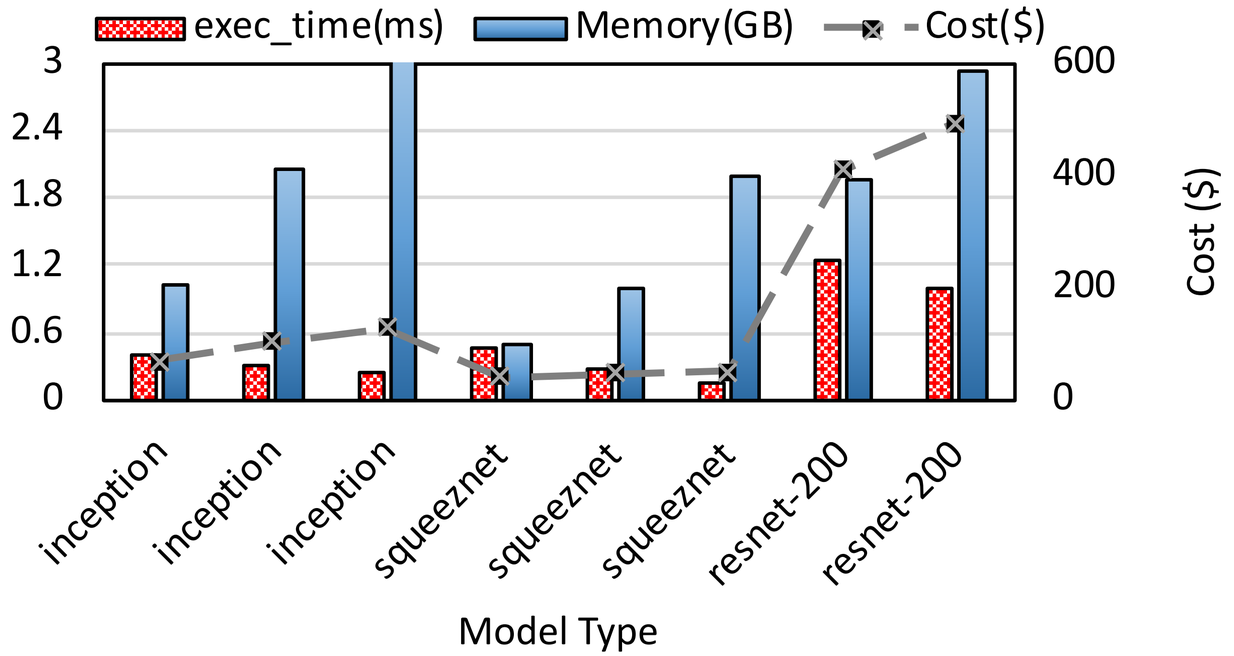
Arrival rate variability



Wiki

Twitter

Serverless Function Configuration



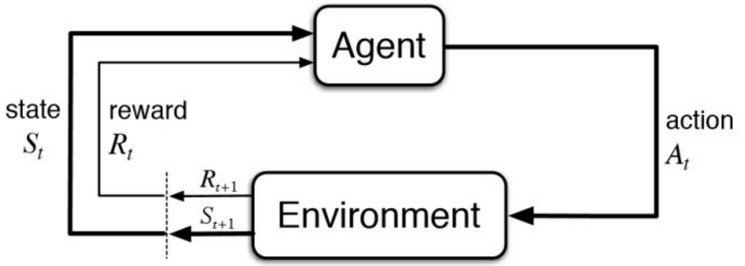
Cost is **1.5x** higher for 0.2x lower latency

What we need?

- How to make the users oblivious of model selection from the extensive pool of models?
- How to right-size VMs and appropriately configure the serverless functions?
- What is the right degree to combine serverless functions along with VMs for dynamic load?

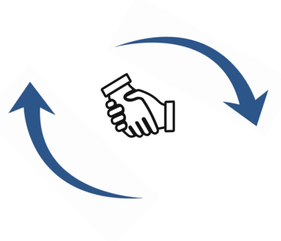
Proposed Solutions

- Feedback-driven learning based model selection.
- Load-Based Procurement Policies
- Provisioning latency and SLO aware resource selection
- Dynamic serverless configurations.



Static Load

Dynamic Load



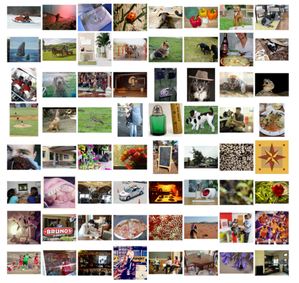
CPU

Memory

Exec Time

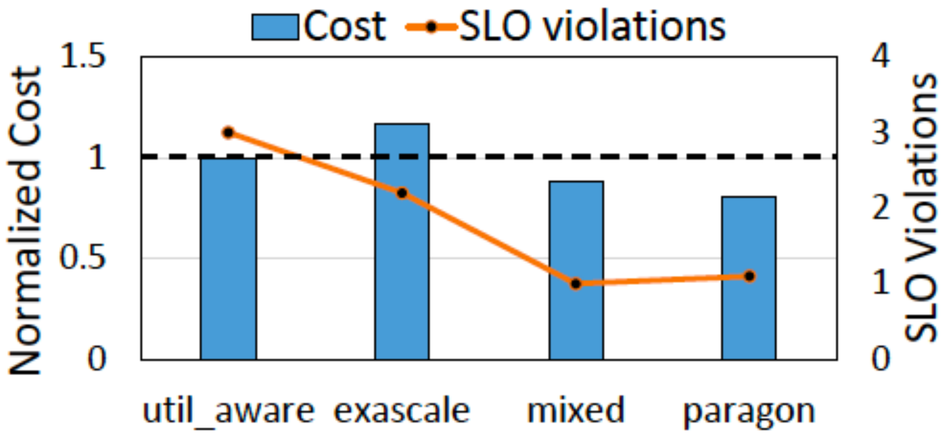
Implementation and Evaluation

- Mxnet Framework.
- AWS resources.
- Pretrained ML models on imagenet dataset.

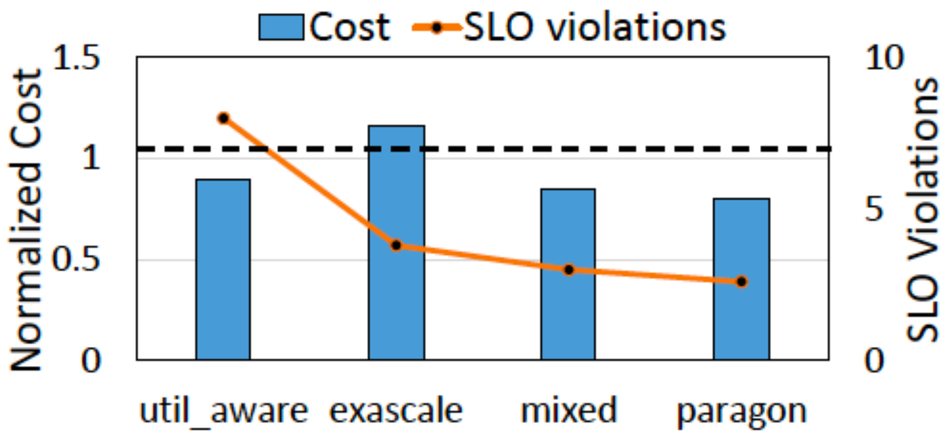


Query Type	Memory Required (GB)	Memory Allocated (GB)	Average Execution (ms)	Requests per vCPU for VMs
Caffenet	1.024	3.072	300	4
Googlenet	0.456	2.048	450	3
Squeezenet	0.154	2.048	130	6
Resnet-18	0.304	3.072	320	3
Resnet-200	1.024	3.072	956	1
Resnext-50	0.645	3.072	560	2

Initial Results



(a) Workload-1: Berkeley Trace.



(b) Workload-1: WITS Trace.

60% less SLO Violations.
10% reduction in deployment costs


Q&A
questions
answers
asking
investigation
inquiry
info
ask
query
q&a
help
problem
confused
investigation
inquiry
answers
questions
help
problem
ask
answers
inquiry
info
ask
inquiry
q & a
answers
help
investigation
ask
problem
confused
asking
query
problem
info
inquiry
questions
help
info
help
q&a
help
inquiry
answers
investigation
query
help
questions
problem
asking
confused
info
inquiry
questions
help
ask
answers
inquiry
help
investigation
inquiry
rs
fused