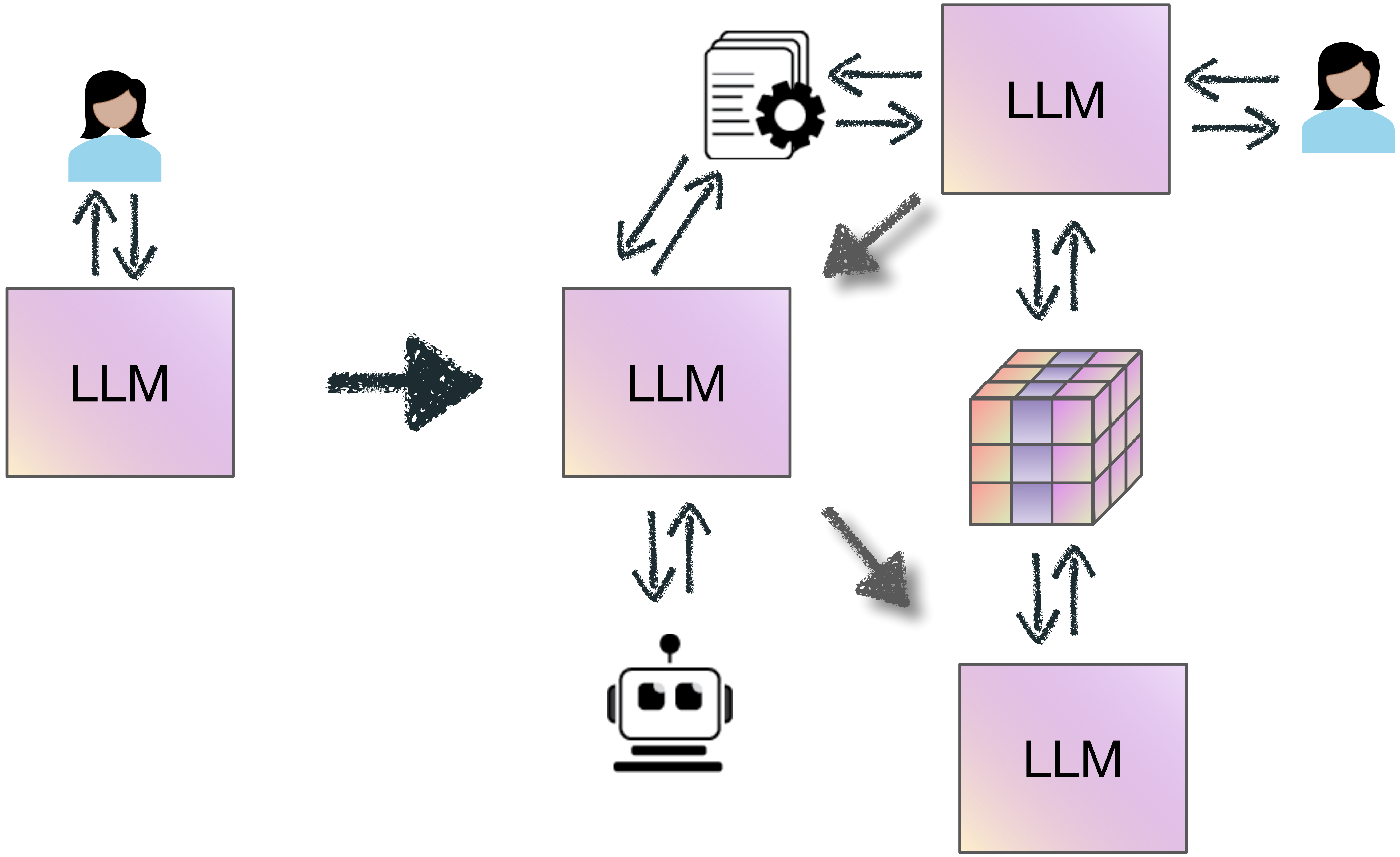
Reconciling Accuracy, Cost, and Latency of Inference Serving Systems

Pooyan Jamshidi

University of South Carolina

https://pooyanjamshidi.github.io/

Modern ML Systems are increasingly composed of multiple interdependent components



Performance tradeoff in **ML pipelines** becomes orders of magnitude more difficult than the single-task ML systems

- **Compositional Complexity**: Dependencies propagate performance impacts.
- **Cascading Tradeoffs**: Latency and cost propagate through the graph. Accuracy depends on intermediate representations.
- **Tradeoff Analysis:** Improving one node may degrade overall performance.
- **Optimization** often requires a **joint search** over an exponentially large search space!

ML inference services have **strict** requirements

Highly Responsive!

ML inference services have **strict** requirements

Highly Responsive!

Cost-Efficient!

ML inference services have **strict** requirements

Highly Responsive!

Cost-Efficient!

Highly Accurate!

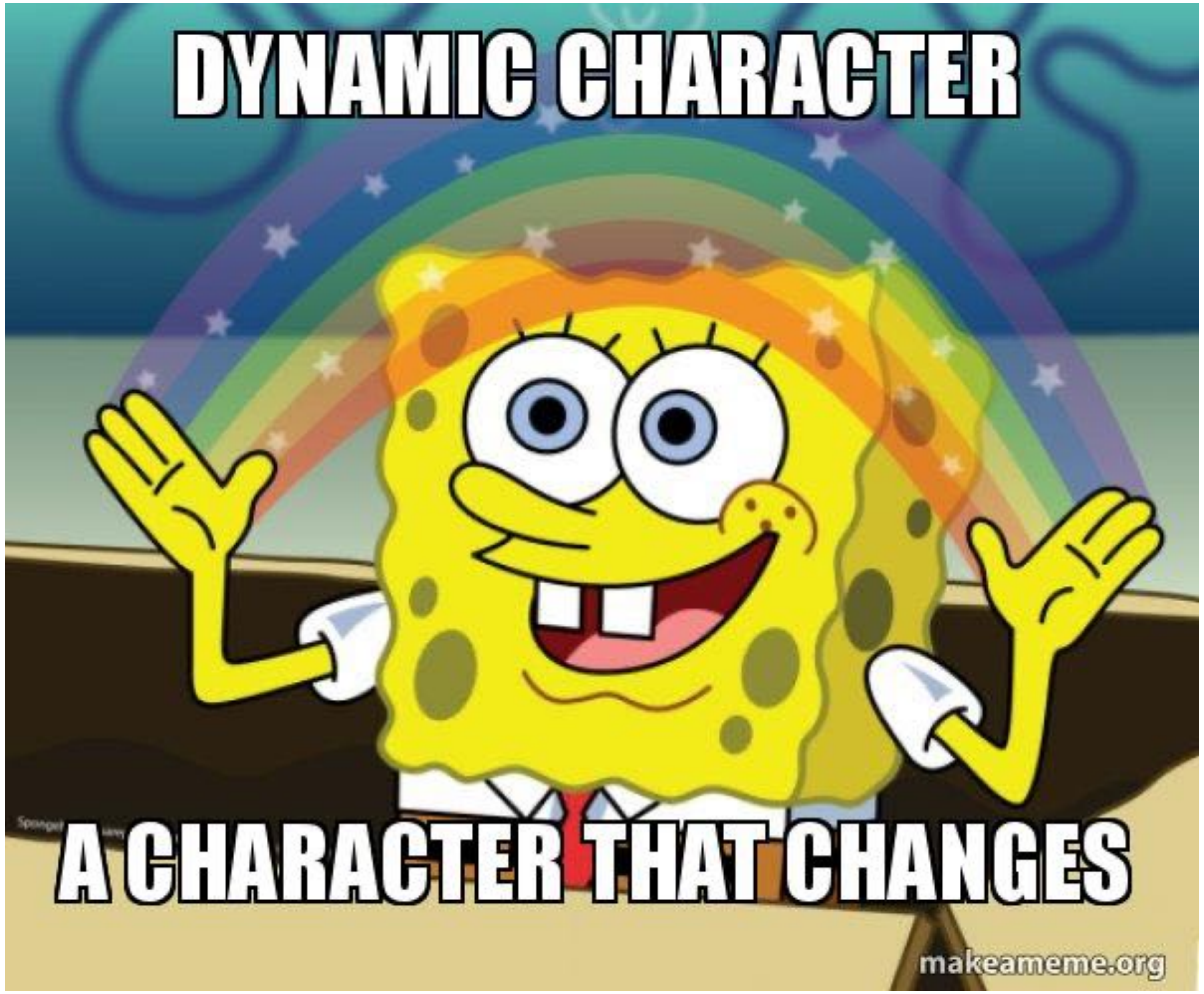
ML inference services have strict **& conflicting** requirements

Highly Responsive!

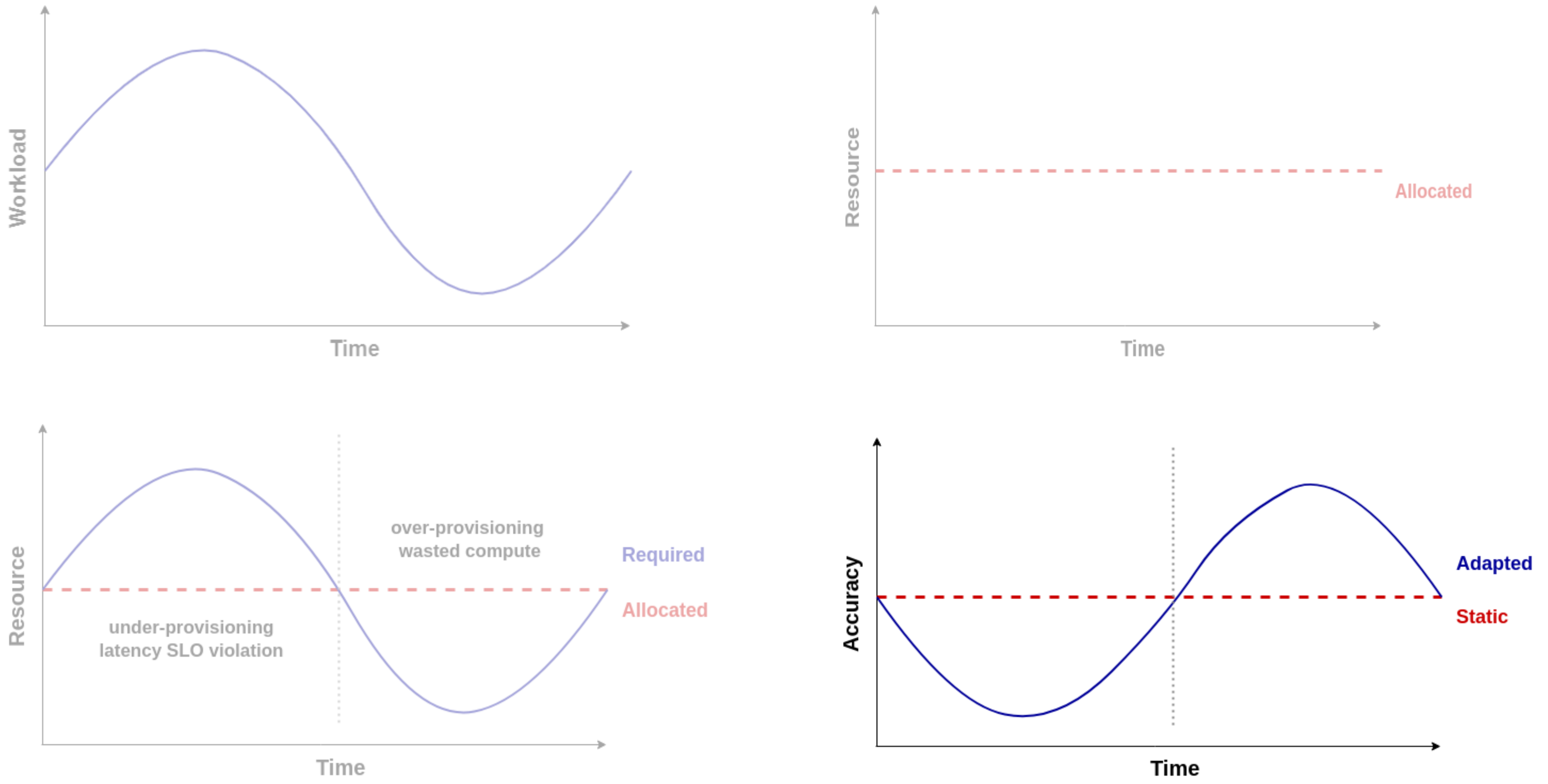
Cost-Efficient!

Highly Accurate!

More challenge: **Dynamic** workload



Quality adaptation

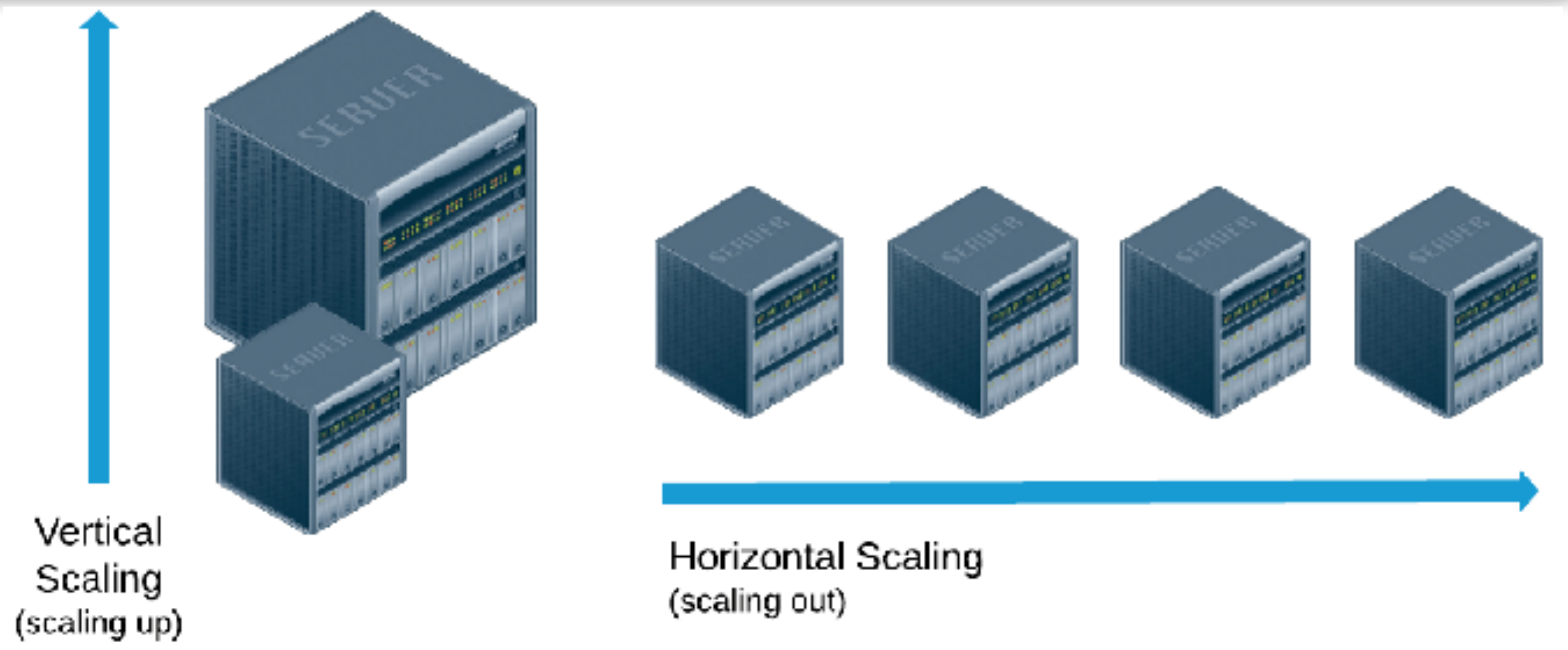


Our work, similar to all other research publications, stands on the shoulders of giants :)

Resource Scaling

Vertical Scaling (AutoPilot EuroSys’20)

Horizontal Scaling (MArk ATC’19)

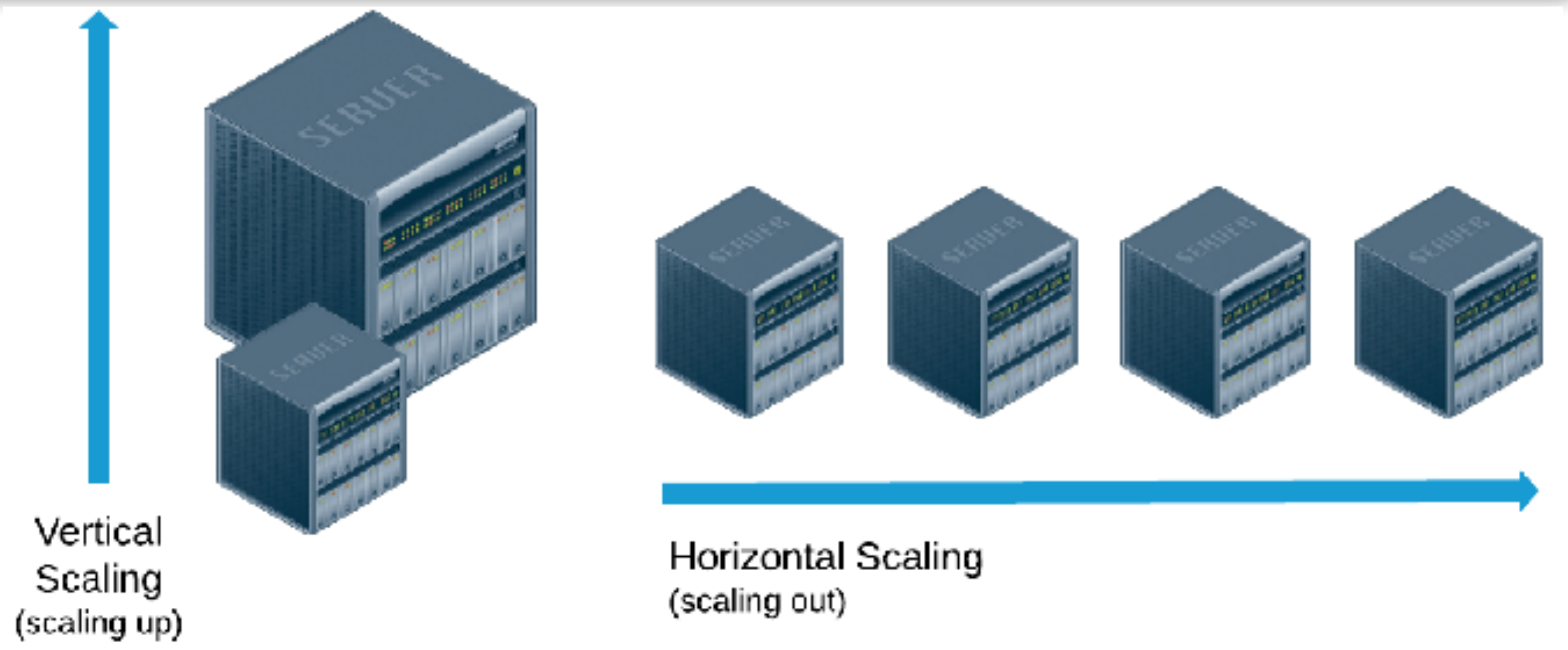


Our work, similar to all other research publications, stands on the shoulders of giants :)

Resource Scaling

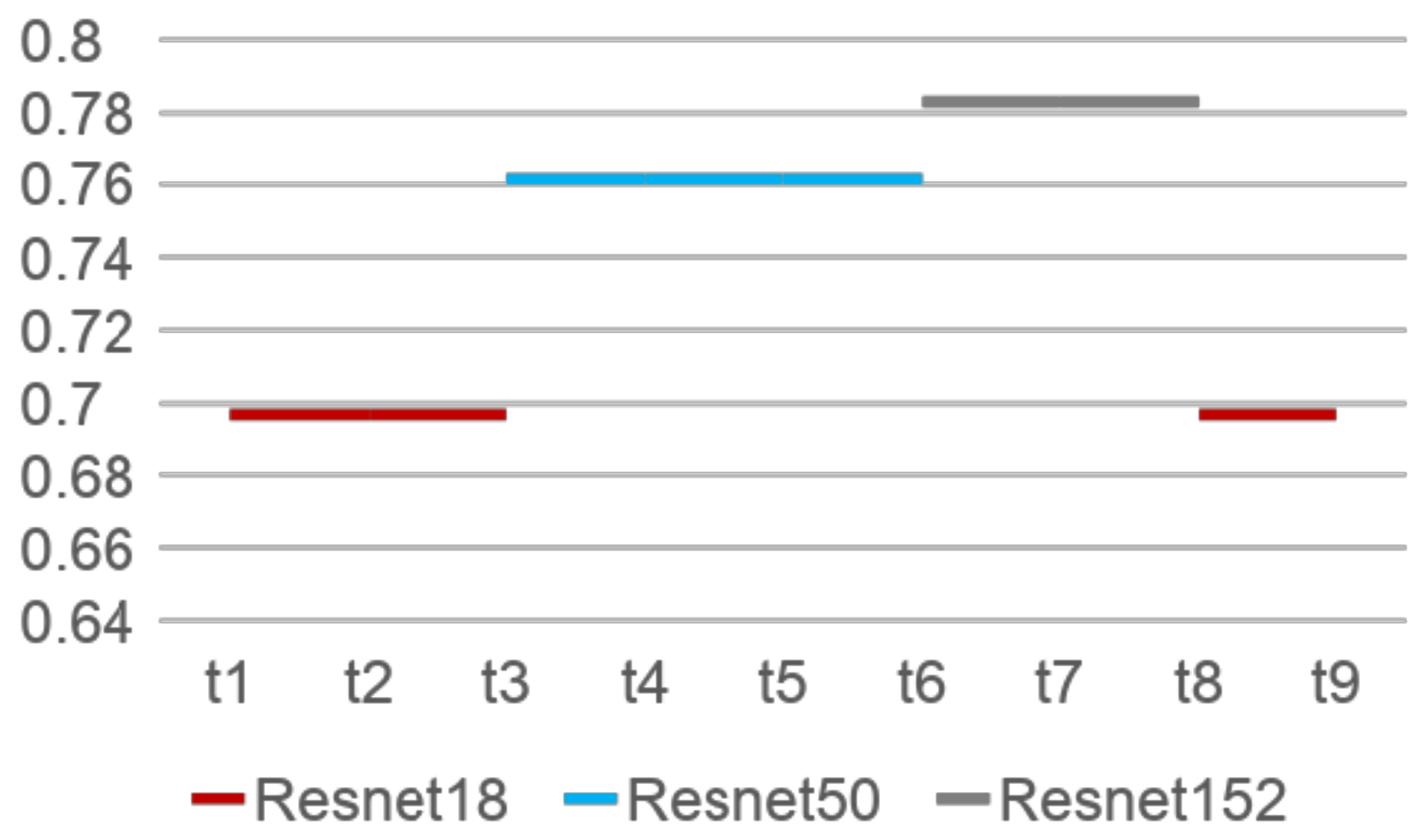
Vertical Scaling (AutoPilot EuroSys'20)

Horizontal Scaling (MArk ATC'19)



Quality Adaptation

Model Variants (Model-Switching Hotcloud'20)



Solutions Preview:

\\ InfAdapter, IPA, and Sponge

EuroMLSys

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

Mehran Salmani[*], Saeid Ghafouri[§‡], Alireza Sanaee[§], Kamran Razavi[†], Max Mühlhäuser[†], Joseph Doyle[§], Pooyan Jamshidi[‡], Mohsen Sharifi[*]

Iran University of Science and Technology[*], Queen Mary University of London[§], Technical University of Darmstadt[†], University of South Carolina[‡]

JSys Journal of Systems Research Volume 4, Issue 1, April 2024

[SOLUTION] IPA: INFERENCE PIPELINE ADAPTATION TO ACHIEVE HIGH ACCURACY AND COST-EFFICIENCY

Saeid Ghafouri
University of South Carolina & Queen Mary University of London

Kamran Razavi
Technical University of Darmstadt

Mehran Salmani
Technical University of Ilmenau

Alireza Sanaee
Queen Mary University of London

Tania Lorido Botran
Roblox

Lin Wang
Paderborn University

Joseph Doyle
Queen Mary University of London

Pooyan Jamshidi
University of South Carolina

EuroMLSys

Sponge: Inference Serving with Dynamic SLOs Using In-Place Vertical Scaling

Kamran Razavi[*]
Technical University of Darmstadt

Saeid Ghafouri[*]
Queen Mary University of London

Max Mühlhäuser
Technical University of Darmstadt

Pooyan Jamshidi
University of South Carolina

Lin Wang
Paderborn University

Problem:

Multi-Objective Optimization with Known Constraints under Uncertainty

$$
\begin{aligned}
\max \quad & \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC) \\
\text{subject to} \quad & \lambda \leq \sum_{m \in M} th_m(n_m), \\
& \lambda_m \leq th_m(n_m) \\
& p_m(n_m) \leq L, \forall m \in M, \\
& RC \leq B,
\end{aligned}
$$

Solutions:

Different Assumptions

InfAdapter [2023]:
Autoscaling for
ML Inference

IPA [2024]:
Autoscaling for
ML Inference Pipeline

Sponge [2024]:
Autoscaling for
ML Inference Pipeline
Dynamic SLO

EuroMLSys

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

Mehran Salmani*, Saeid Ghafouri§‡, Alireza Sanaee§, Kamran Razavi†, Max Mühlhäuser†, Joseph Doyle§, Pooyan Jamshidi‡, Mohsen Sharifi*

Iran University of Science and Technology*, Queen Mary University of London§, Technical University of Darmstadt†, University of South Carolina‡

InfAdapter [2023]: Autoscaling for ML **Model Inference**

JSys Journal of Systems Research Volume 4, Issue 1, April 2024

[SOLUTION] IPA: INFERENCE PIPELINE ADAPTATION TO ACHIEVE HIGH ACCURACY AND COST-EFFICIENCY

Saeid Ghafouri, *University of South Carolina & Queen Mary University of London*
Kamran Razavi, *Technical University of Darmstadt*
Mehran Salmani, *Technical University of Ilmenau*
Alireza Sanaee, *Queen Mary University of London*
Tania Lorido Botran, *Roblox*
Lin Wang, *Paderborn University*
Joseph Doyle, *Queen Mary University of London*
Pooyan Jamshidi, *University of South Carolina*

IPA [2024]: Autoscaling for **ML Inference Pipeline**

EuroMLSys

Sponge: Inference Serving with Dynamic SLOs Using In-Place Vertical Scaling

Kamran Razavi*, Technical University of Darmstadt
Saeid Ghafouri*, Queen Mary University of London
Max Mühlhäuser, Technical University of Darmstadt
Pooyan Jamshidi, University of South Carolina
Lin Wang, Paderborn University

Sponge [2024]: Autoscaling for ML Inference Pipeline with **Dynamic SLO**

EuroMLSys

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

Mehran Salmani*, Saeid Ghafouri§‡, Alireza Sanaee§, Kamran Razavi†, Max Mühlhäuser†, Joseph Doyle§, Pooyan Jamshidi‡, Mohsen Sharifi*

Iran University of Science and Technology*, Queen Mary University of London§, Technical University of Darmstadt†, University of South Carolina‡

InfAdapter [2023]: Autoscaling for ML **Model Inference**

JSys Journal of Systems Research Volume 4, Issue 1, April 2024

[SOLUTION] IPA: INFERENCE PIPELINE ADAPTATION TO ACHIEVE HIGH ACCURACY AND COST-EFFICIENCY

prior work

IPA [2024]: Autoscaling for **ML Inference Pipeline**

HotCloud '20 ATTEND PROGRAM PARTICIPATE SPONSORS ABOUT

Model-Switching: Dealing with Fluctuating Workloads in Machine-Learning-as-a-Service Systems

Authors:
Jeff Zhang, *New York University;* Sameh Elnikety, *Microsoft Research;* Shuayb Zarar and Atul Gupta, *Microsoft;* Siddharth Garg, *New York University*

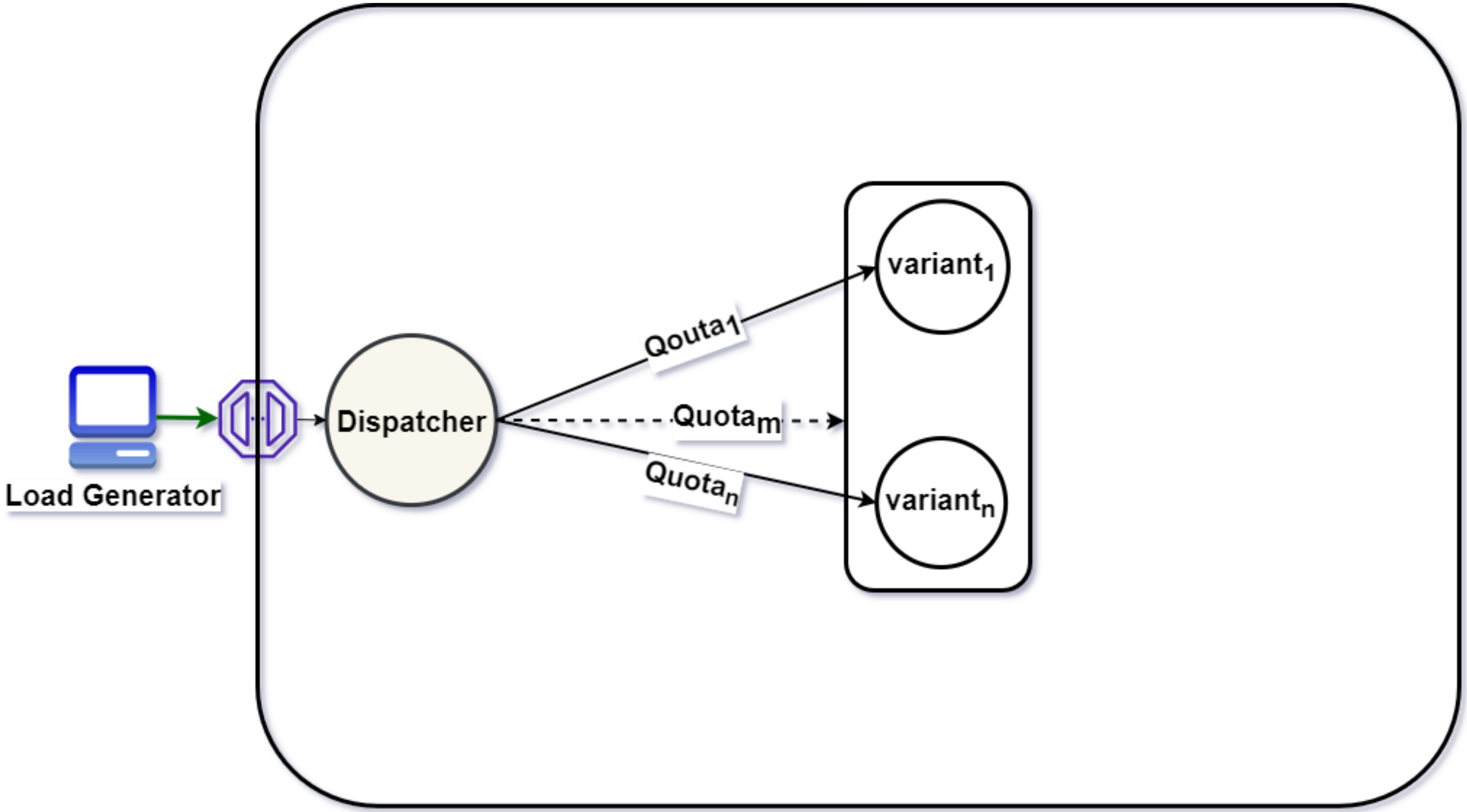
Abstract:
Machine learning (ML) based prediction models, and especially deep neural networks (DNNs) are increasingly being served in the cloud in order to provide fast and accurate inferences. However, existing service ML serving systems have trouble dealing with fluctuating workloads and either drop requests or significantly expand hardware resources in response to load spikes. In this paper, we introduce Model-Switching, a new approach to dealing with fluctuating workloads when serving DNN models. Motivated by the observation that end-users of ML primarily care about the accuracy of responses that are returned within the deadline (which we refer to as effective accuracy), we propose to switch from complex and highly accurate DNN models to simpler but less accurate models in the presence of load spikes. We show that the flexibility introduced by enabling online model switching provides higher effective accuracy in the presence of fluctuating workloads compared to serving using any single model. We implement Model-Switching within Clipper, a state-of-art DNN model serving system, and demonstrate its advantages over baseline approaches.

Pooyan Jamshidi
University of South Carolina

Lin Wang
Paderborn University

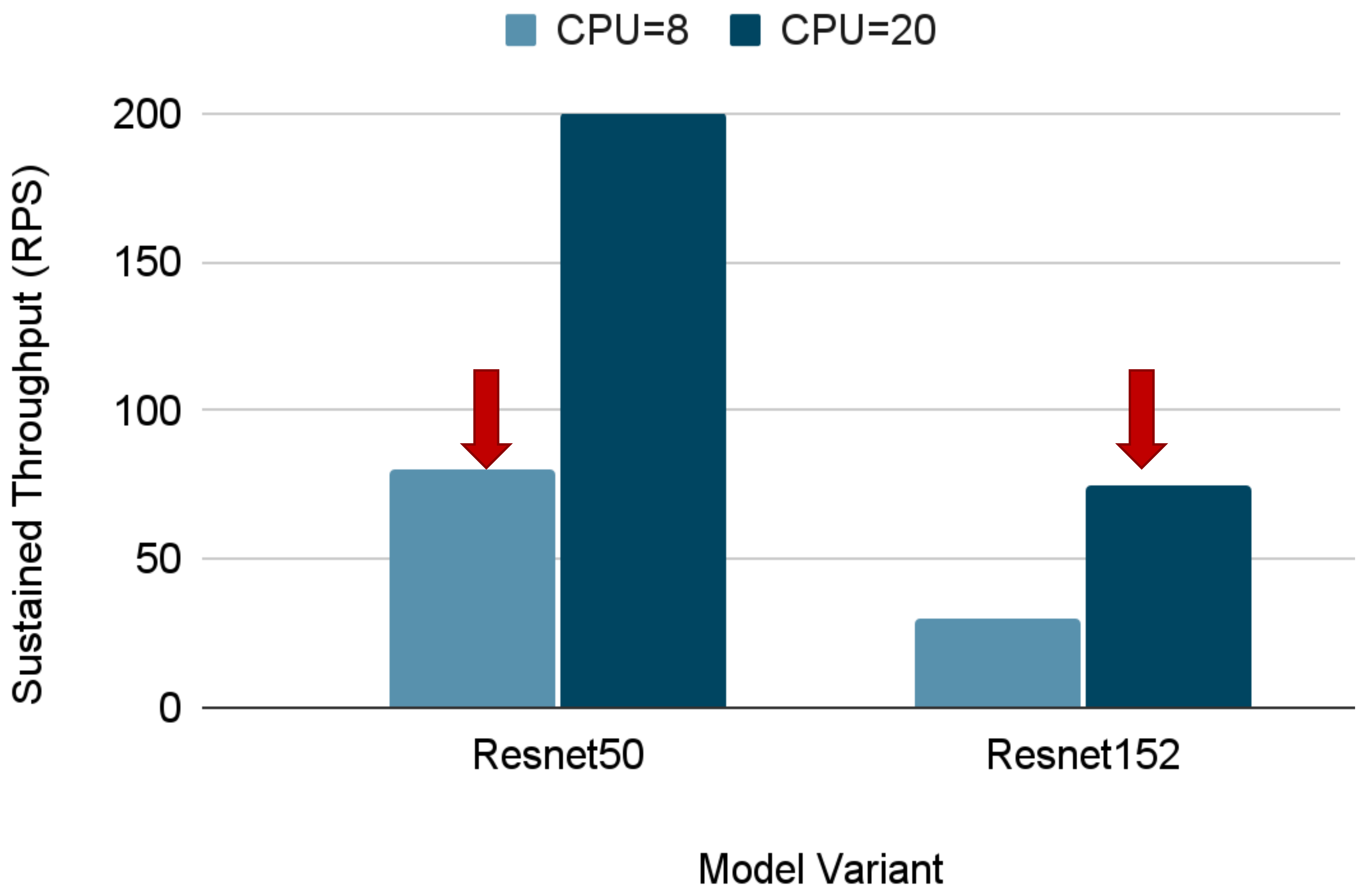
Sponge [2024]: Autoscaling for ML Inference Pipeline with **Dynamic SLO**

InfAdapter (our solution) vs. Model Switching (prior work)



Selecting a **subset of model variants**, each having its size meeting latency requirements for the predicted workload while **maximizing accuracy and minimizing resource cost**

First insight: The same throughput can be achieved with different computing resources by switching the model variants



ResNet-50:

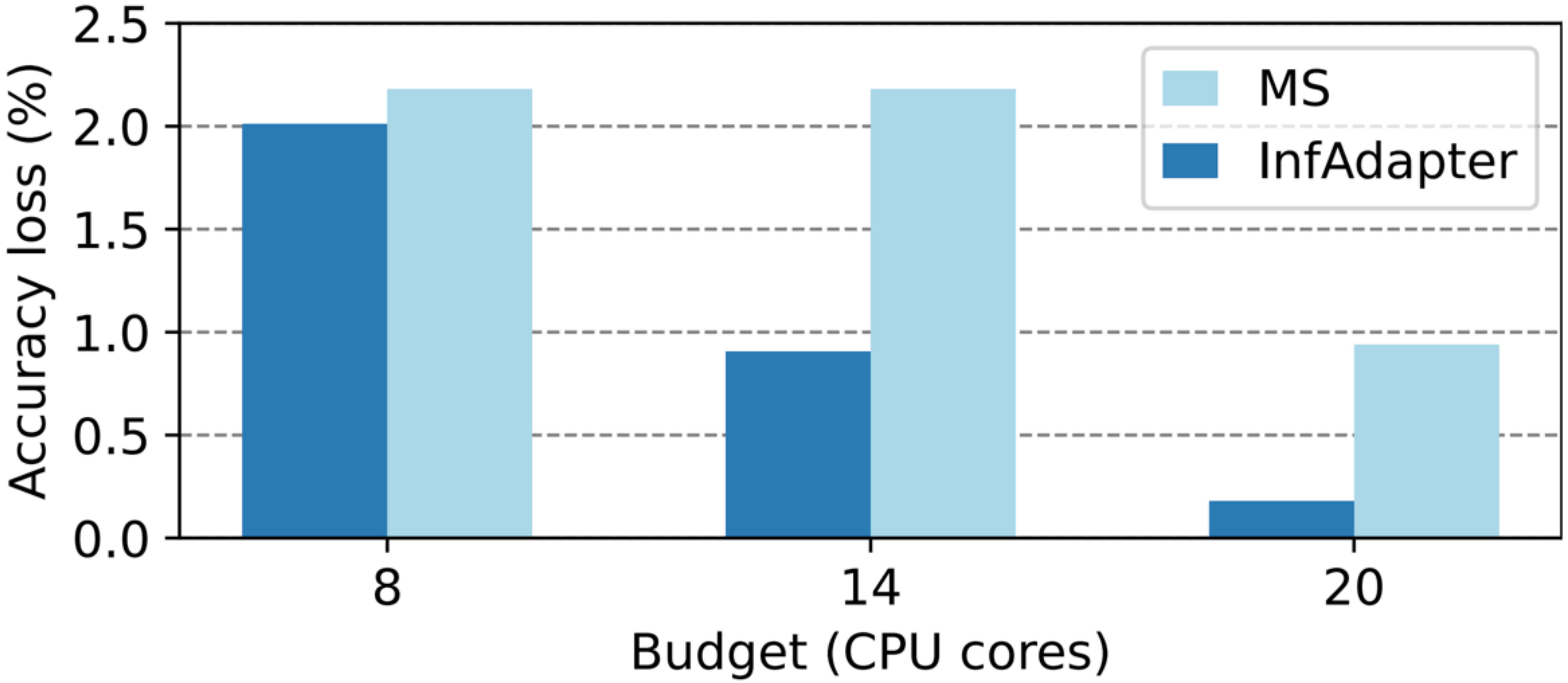
- **Depth**: 50 layers.
- **Top-1 Accuracy**: ~76-77% on ImageNet.
- **Top-5 Accuracy**: ~93-94% on ImageNet.
- **Model Size**: Smaller, faster to train and deploy.

ResNet-152:

- **Depth**: 152 layers.
- **Top-1 Accuracy**: ~78-80% on ImageNet.
- **Top-5 Accuracy**: ~94.5-95% on ImageNet.
- **Model Size**: Larger, higher computational cost,

Multi-models (our solution—InfAdapter) vs single-model (Model-Switching)

Higher average accuracy by using multiple model variants



InfAdapter: Formulation

$$\max \quad \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC)$$

InfAdapter: Formulation

$$\max \quad \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC)$$

Maximizing Average Accuracy

InfAdapter: Formulation

$$\max \quad \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC)$$

Maximizing Average Accuracy

Minimizing Resource and Loading Costs

InfAdapter: Formulation

$$
\begin{aligned}
\max \quad & \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC) \\
\text{subject to} \quad & \lambda \leq \sum_{m \in M} th_m(n_m), \\
& \lambda_m \leq th_m(n_m) \\
& p_m(n_m) \leq L, \forall m \in M, \\
& RC \leq B, \\
& n_m \in \mathbb{W}, \forall m \in M.
\end{aligned}
$$

InfAdapter: Formulation

$$
\begin{aligned}
\max \quad & \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC) \\
\text{subject to} \quad & \lambda \leq \sum_{m \in M} th_m(n_m), \\
& \lambda_m \leq th_m(n_m) \\
& p_m(n_m) \leq L, \forall m \in M, \\
& RC \leq B, \\
& n_m \in \mathbb{W}, \forall m \in M.
\end{aligned}
$$

Supporting incoming workload

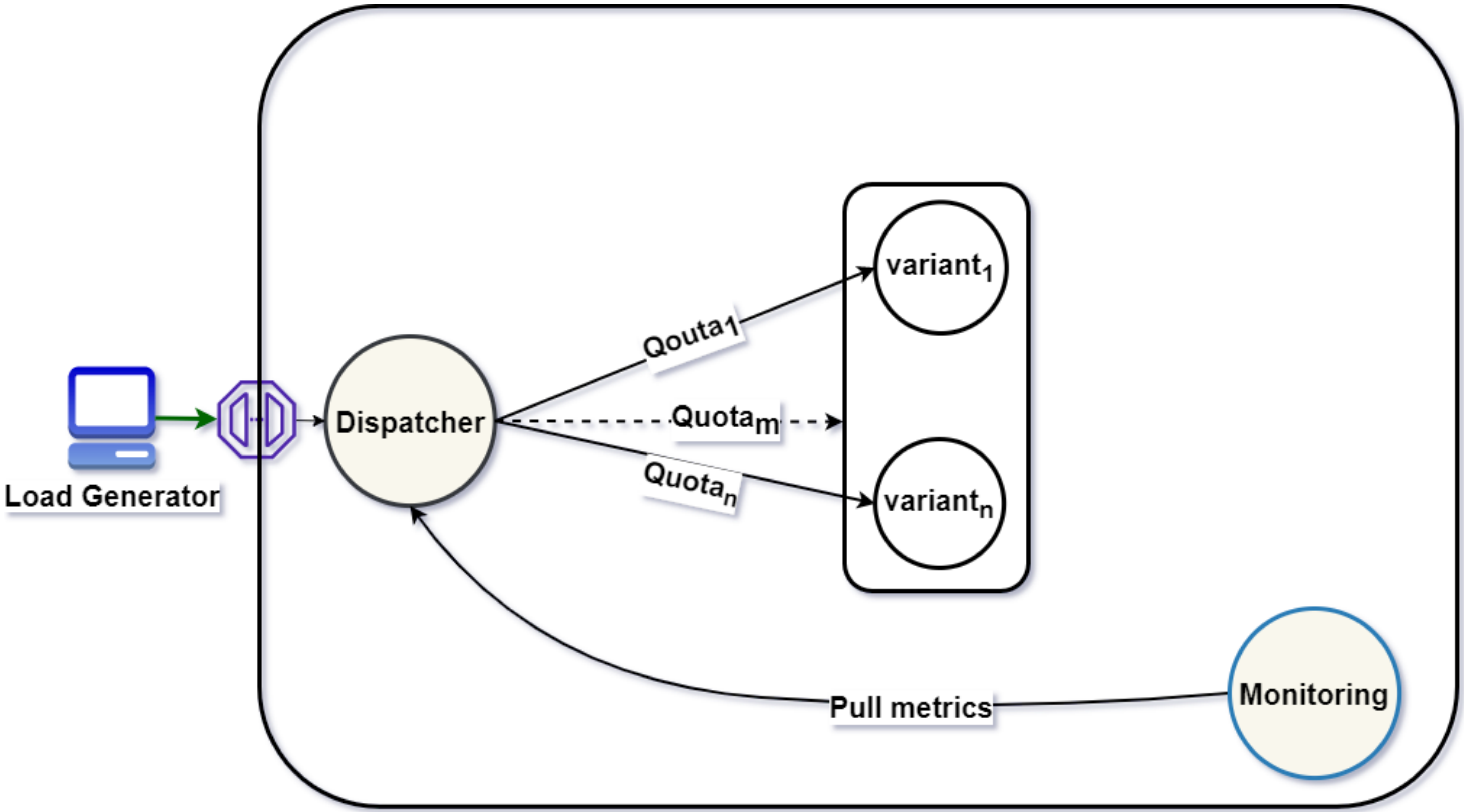
InfAdapter: Formulation

$$
\begin{aligned}
\max \quad & \alpha \cdot AA - (\beta \cdot RC + \gamma \cdot LC) \\
\text{subject to} \quad & \lambda \leq \sum_{m \in M} th_m(n_m), \\
& \lambda_m \leq th_m(n_m) \\
& p_m(n_m) \leq L, \forall m \in M, \\
& RC \leq B, \\
& n_m \in \mathbb{W}, \forall m \in M.
\end{aligned}
$$

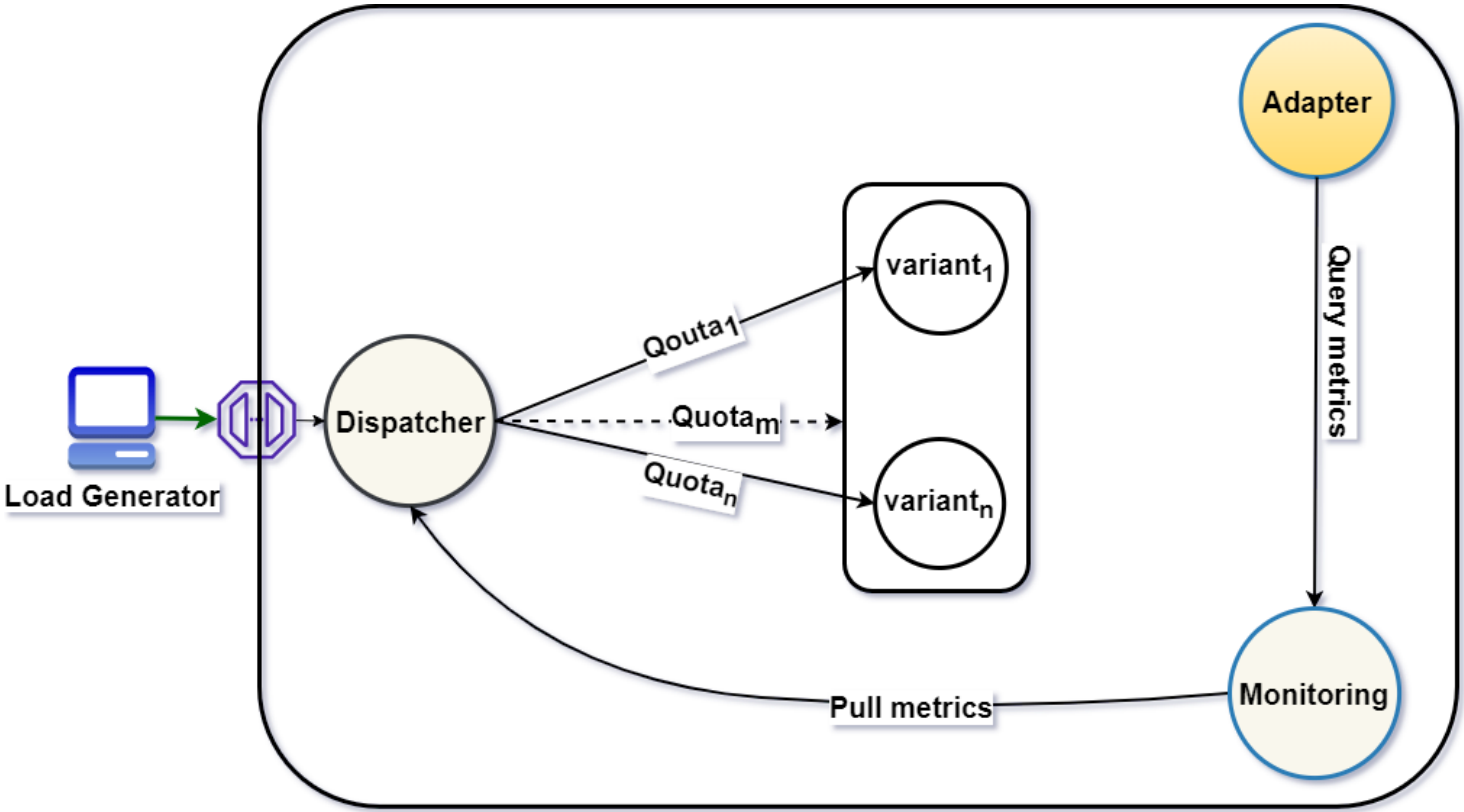
Supporting incoming workload

Guaranteeing end-to-end latency

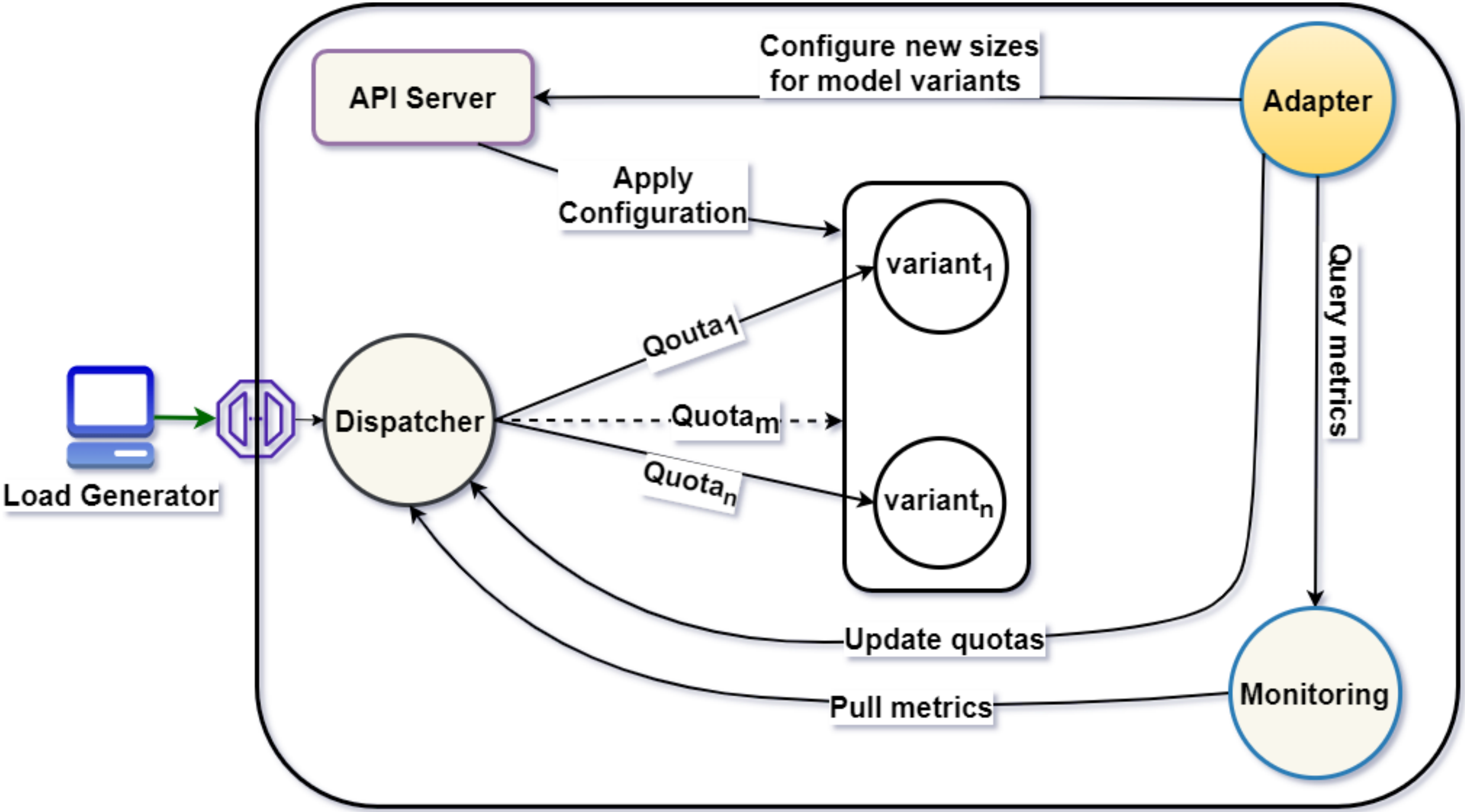
InfAdapter: Design



InfAdapter: Design



InfAdapter: Design



InfAdapter: Experimental evaluation setup

Workload: **Twitter-trace** sample (2022-08)

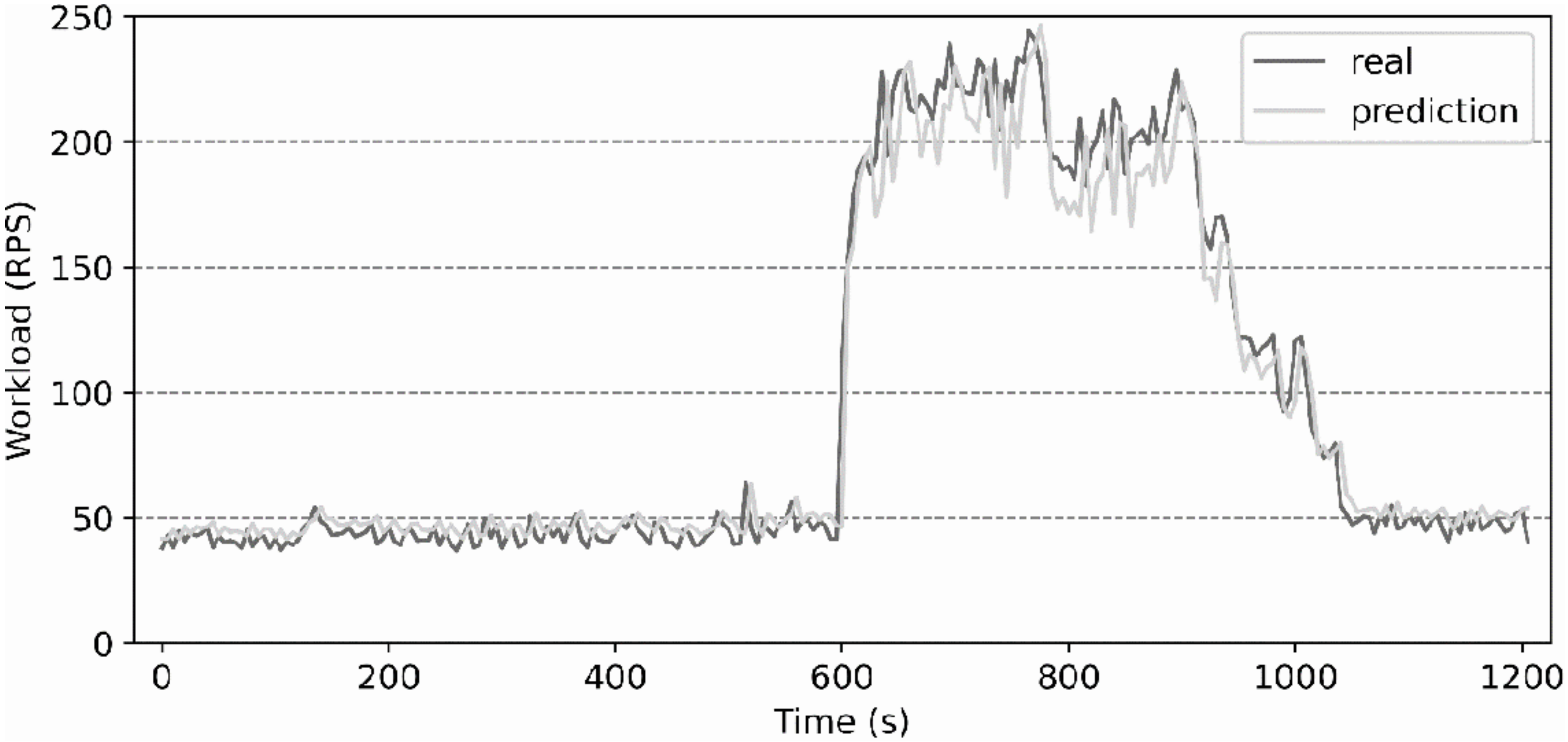
Baselines: **Kubernetes VPA** and **Model-Switching**

Used models: Resnet18, Resnet34, Resnet50, Resnet101, Resnet152

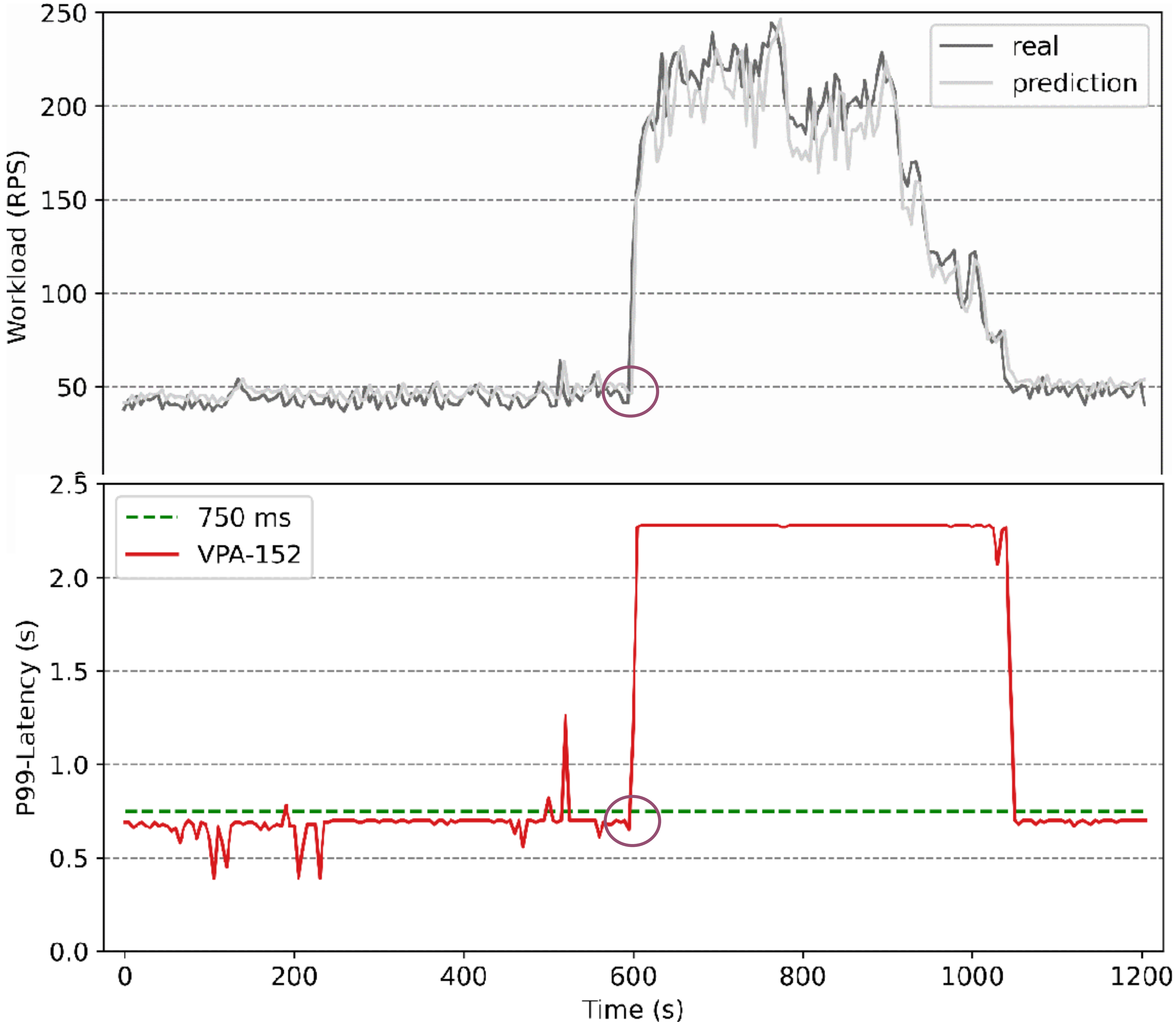
Interval adaptation: 30 seconds

Kubernetes cluster: 48 Cores, 192 GiB RAM

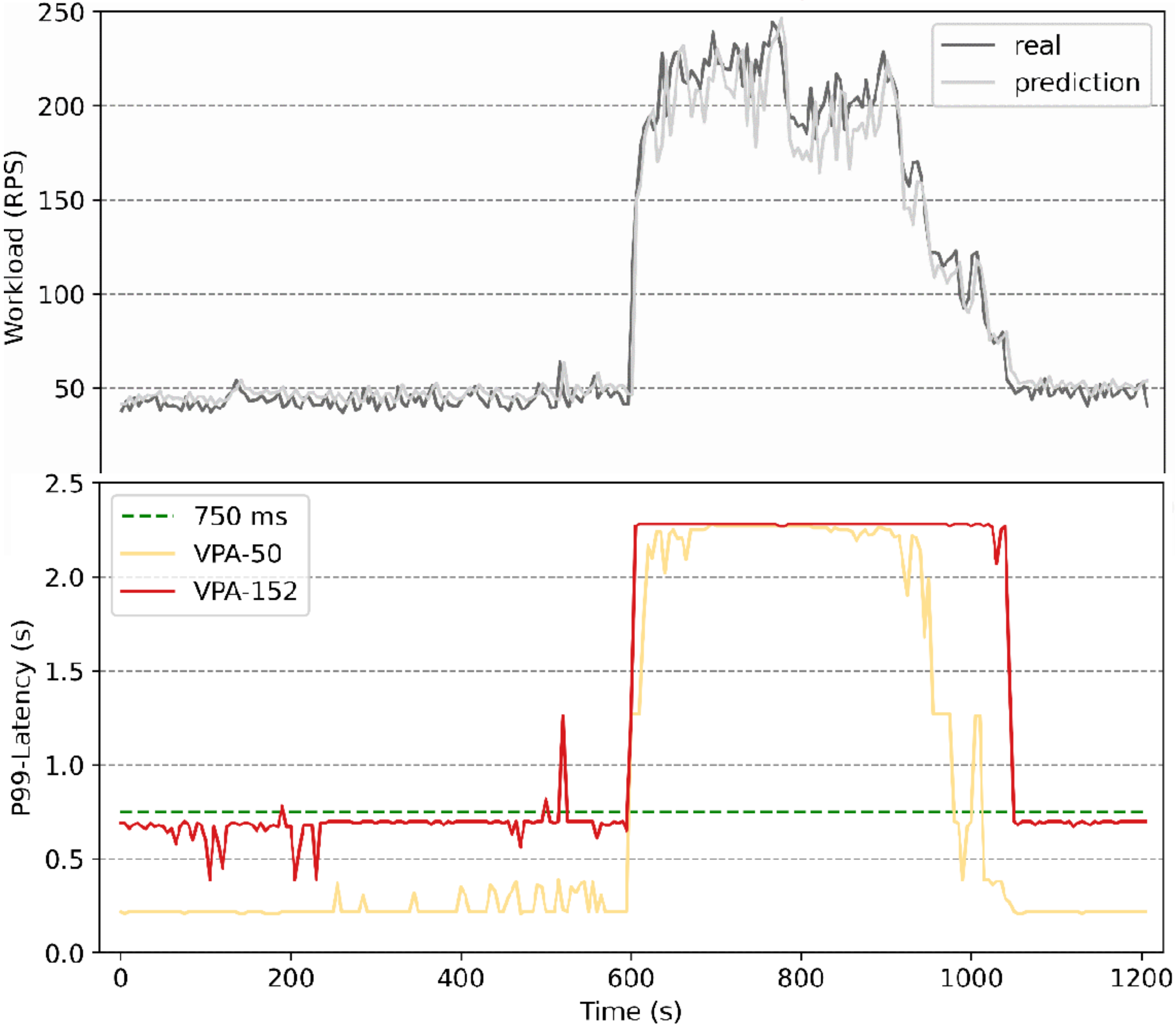
Workload Pattern



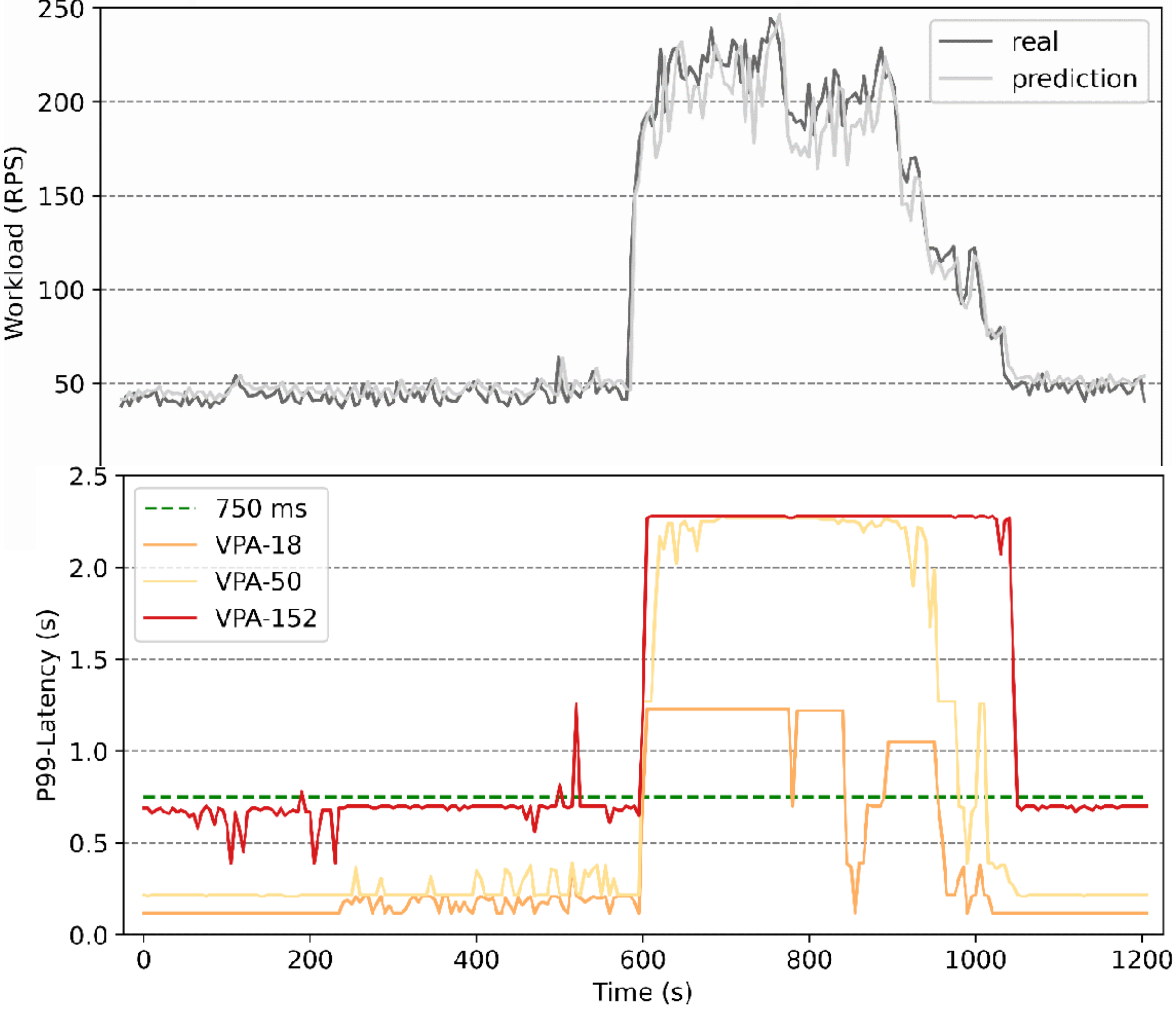
InfAdapter: P99-Latency evaluation



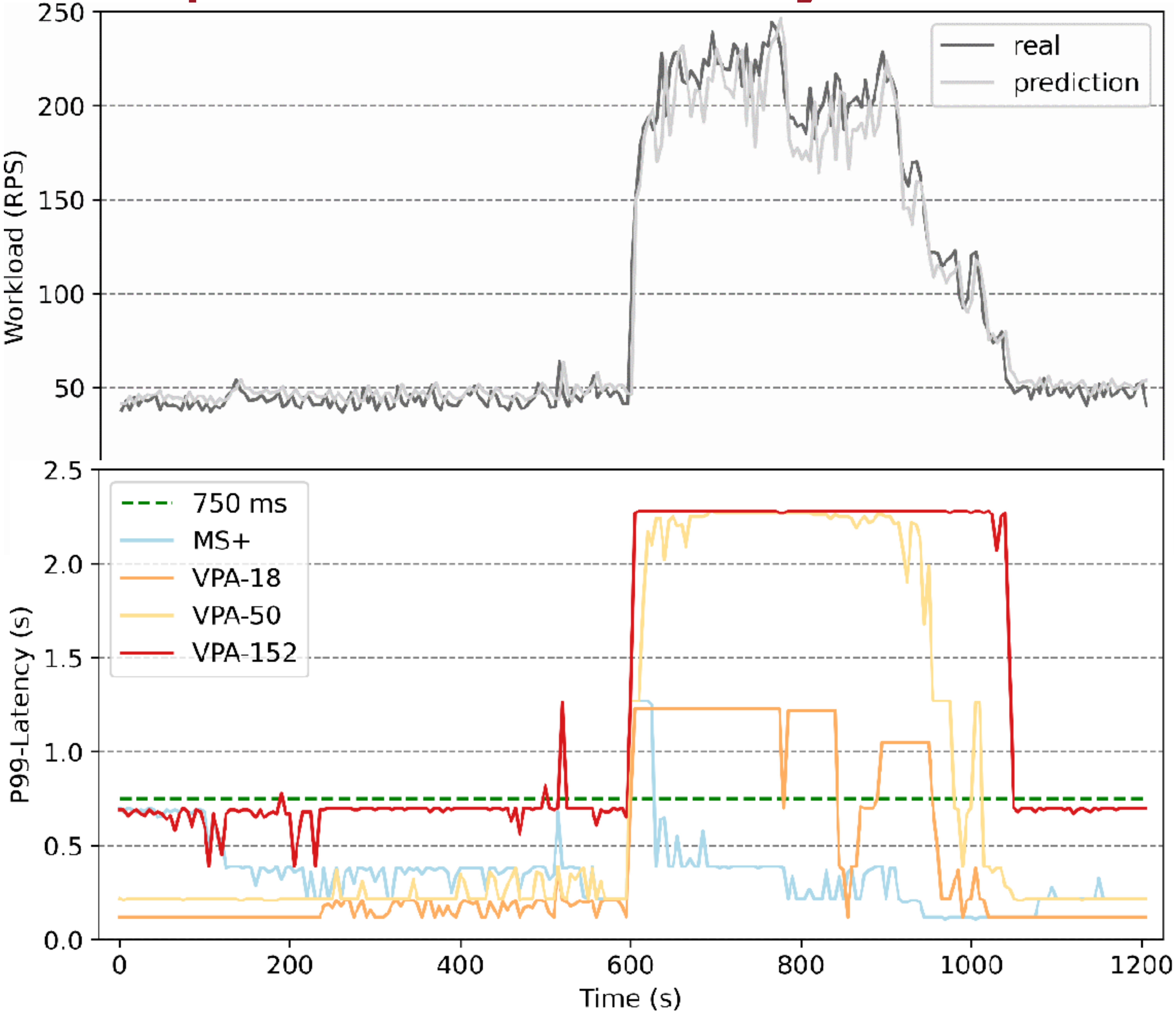
InfAdapter: P99-Latency evaluation



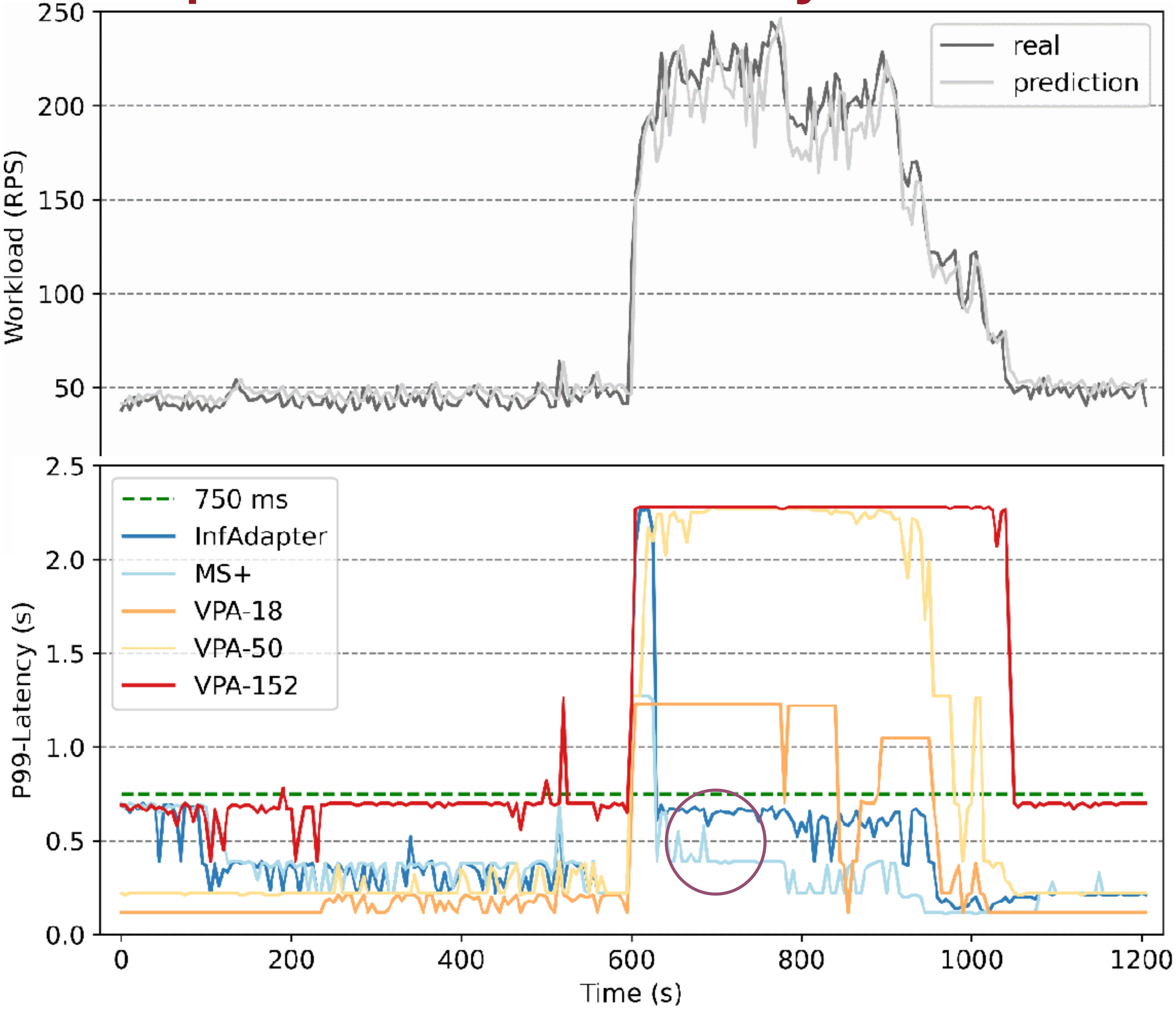
InfAdapter: P99-Latency evaluation



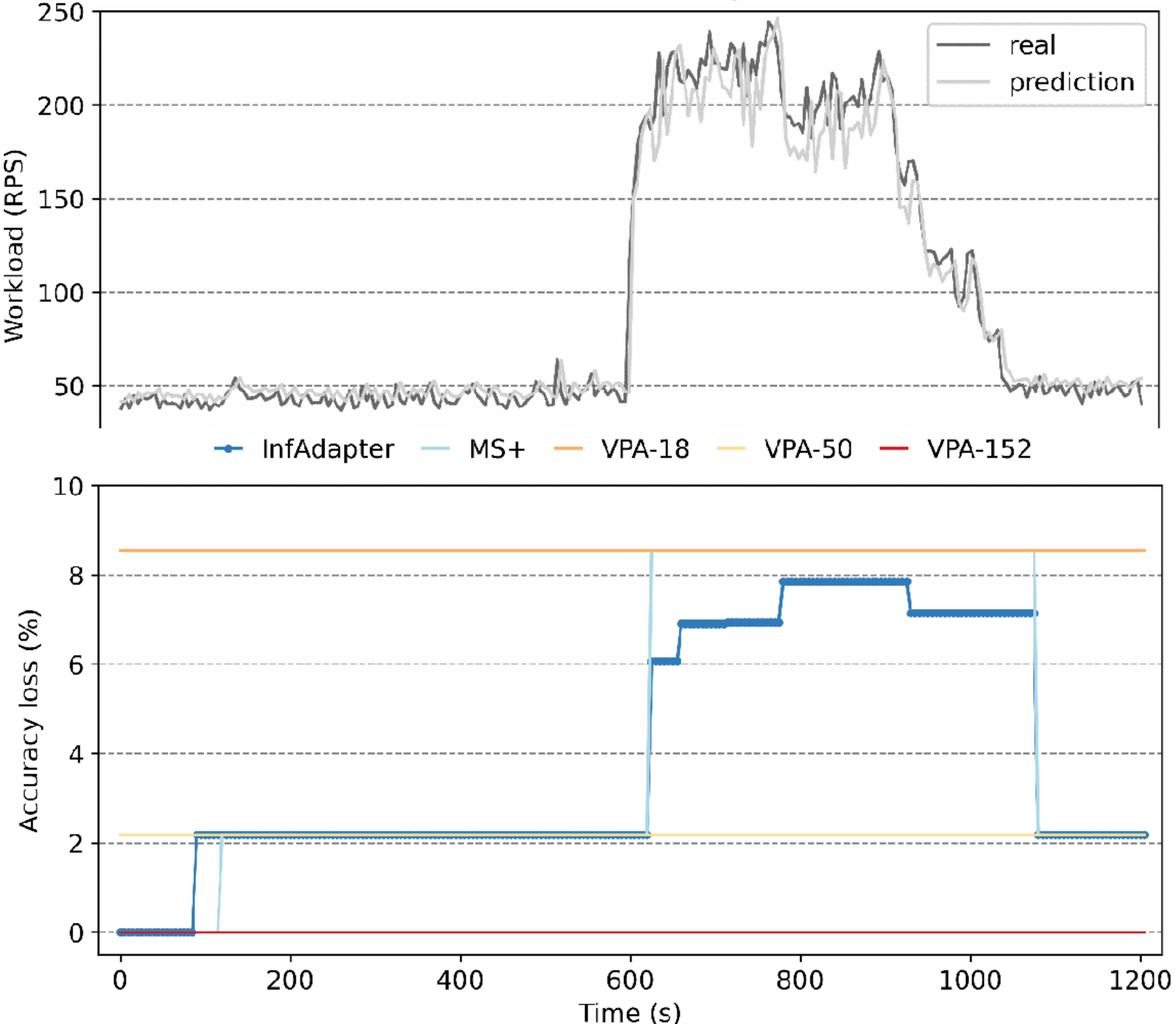
InfAdapter: P99-Latency evaluation



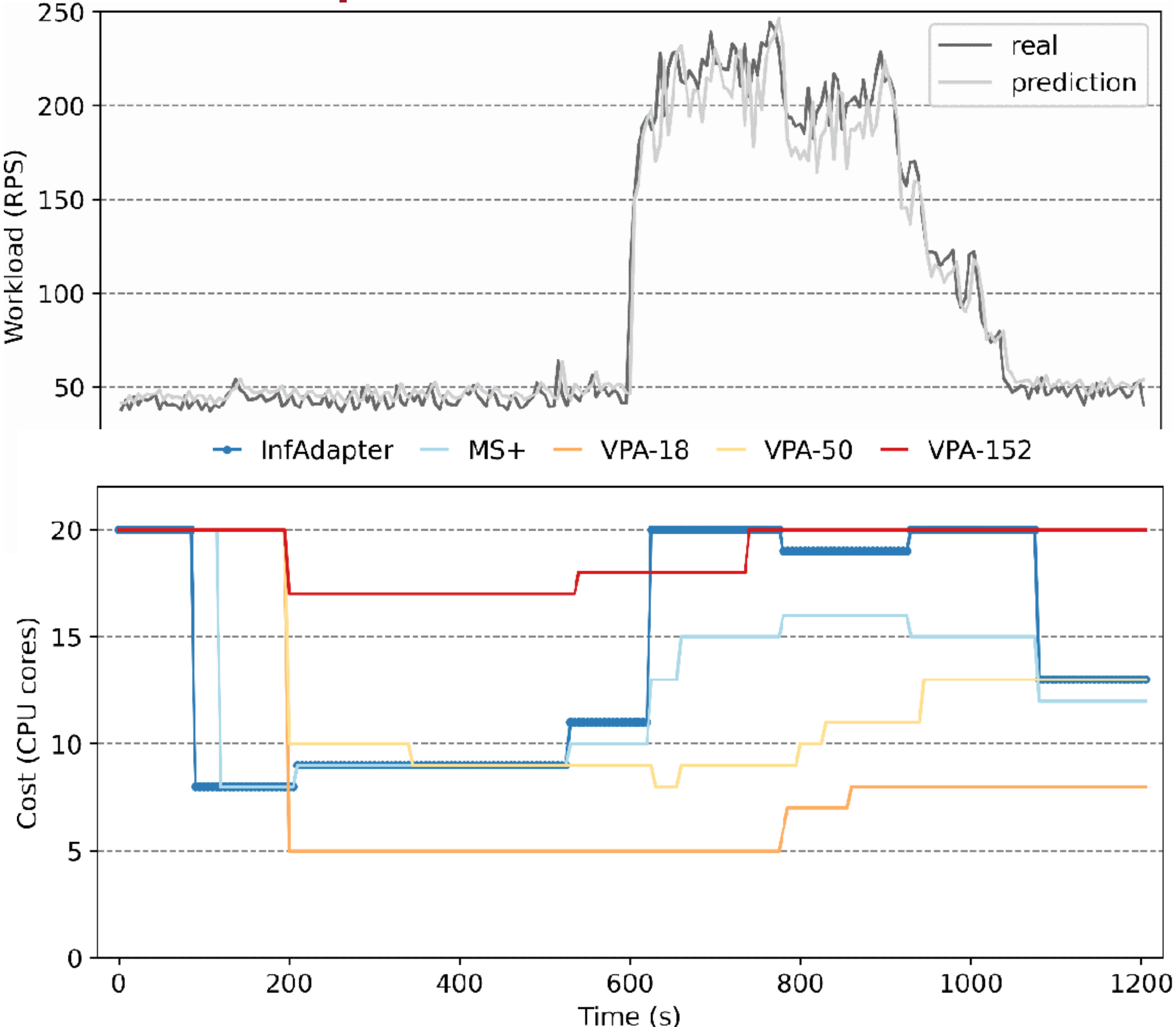
InfAdapter: P99-Latency evaluation



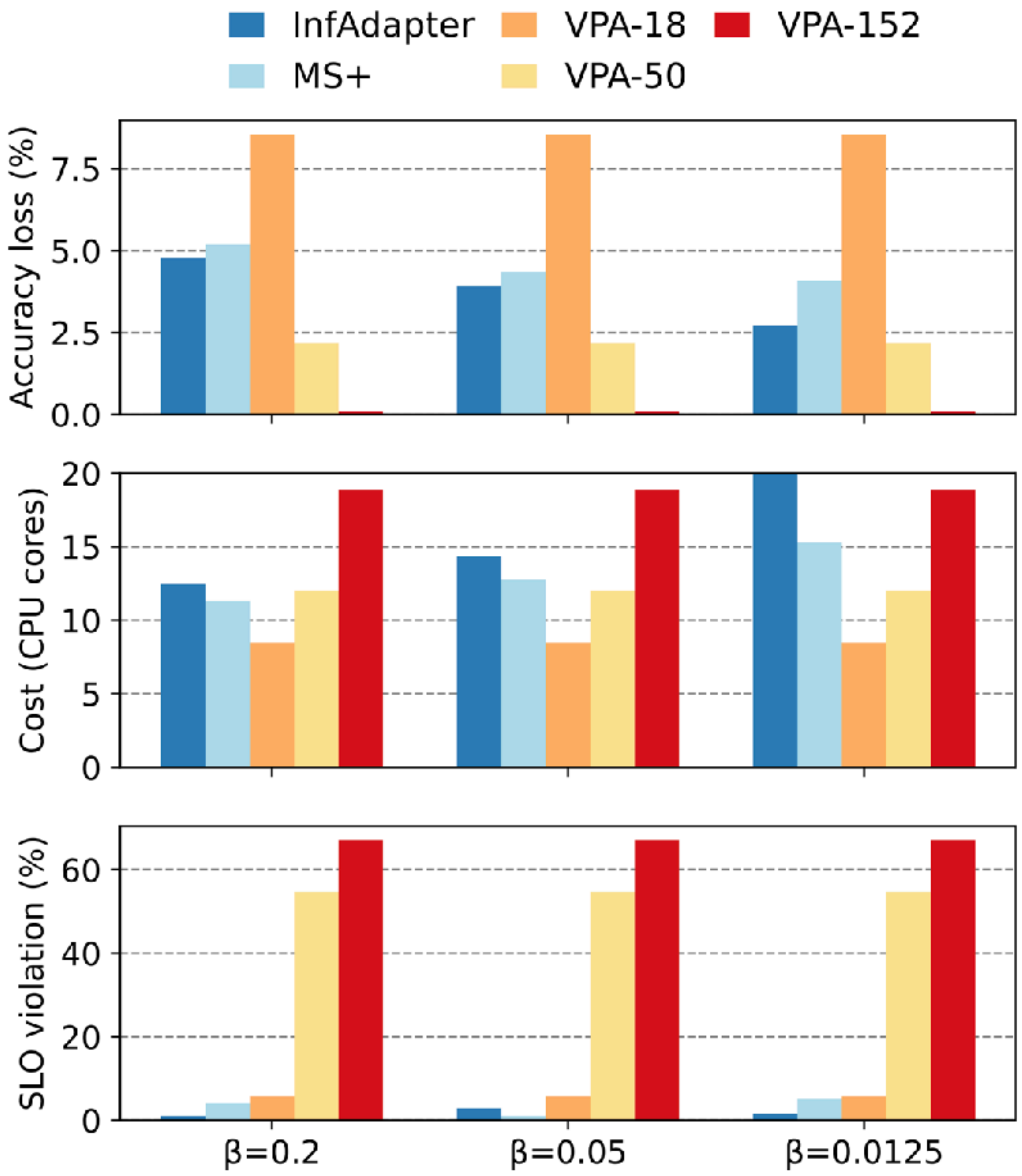
InfAdapter: Accuracy evaluation



InfAdapter: Cost evaluation



InfAdapter: Tradeoff Space



Takeaway

Inference Serving Systems should consider accuracy, latency, and cost at the same time.

Takeaway

Inference Serving Systems should consider accuracy, latency, and cost at the same time.

Model variants provide the opportunity to reduce resource costs while adapting to the dynamic workload.

Using a set of model variants simultaneously provides higher average accuracy compared to having one variant.

Takeaway

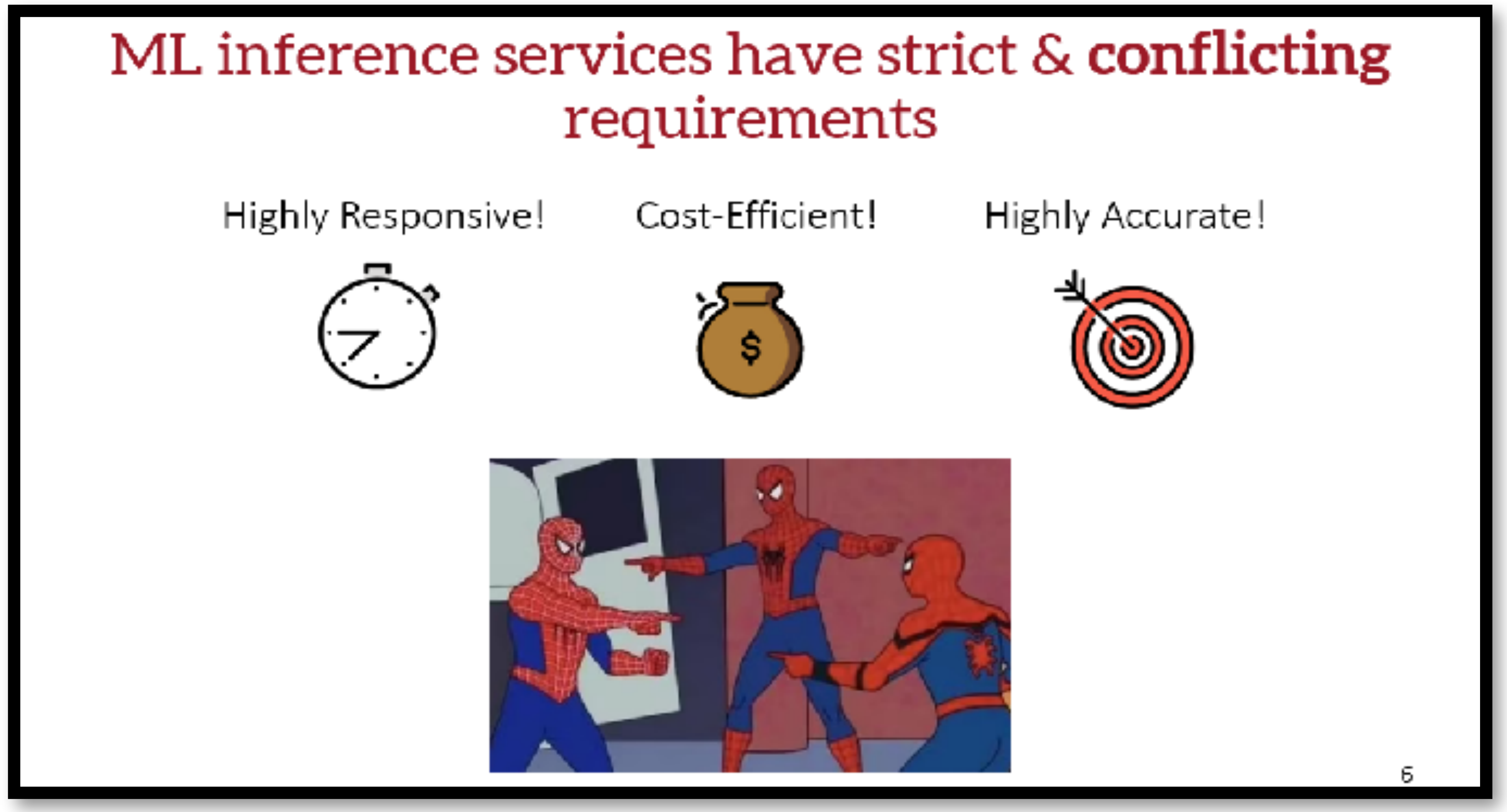
Inference Serving Systems should consider accuracy, latency, and cost at the same time.

Model variants provide the opportunity to reduce resource costs while adapting to the dynamic workload.

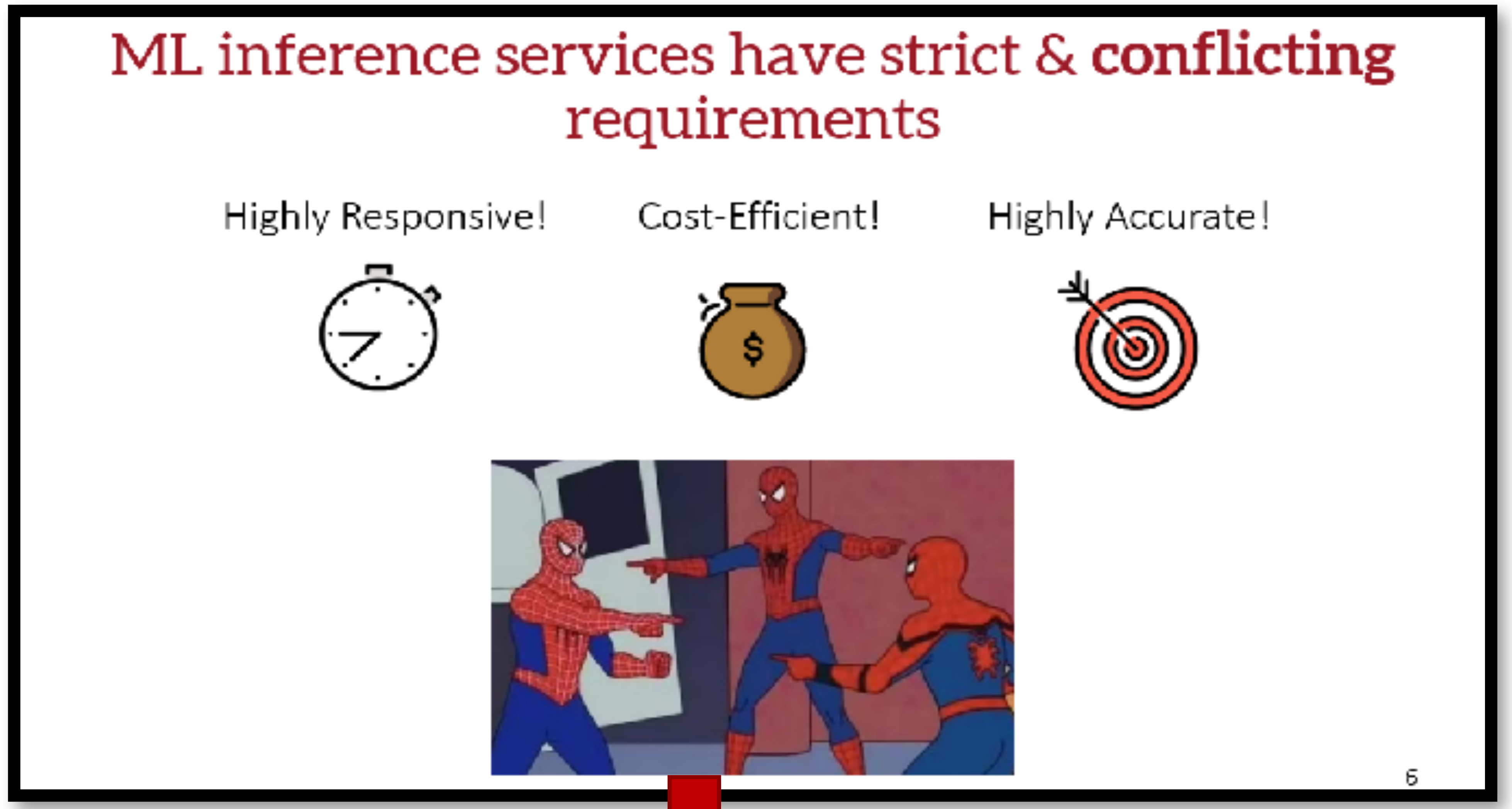
Using a set of model variants simultaneously provides higher average accuracy compared to having one variant.

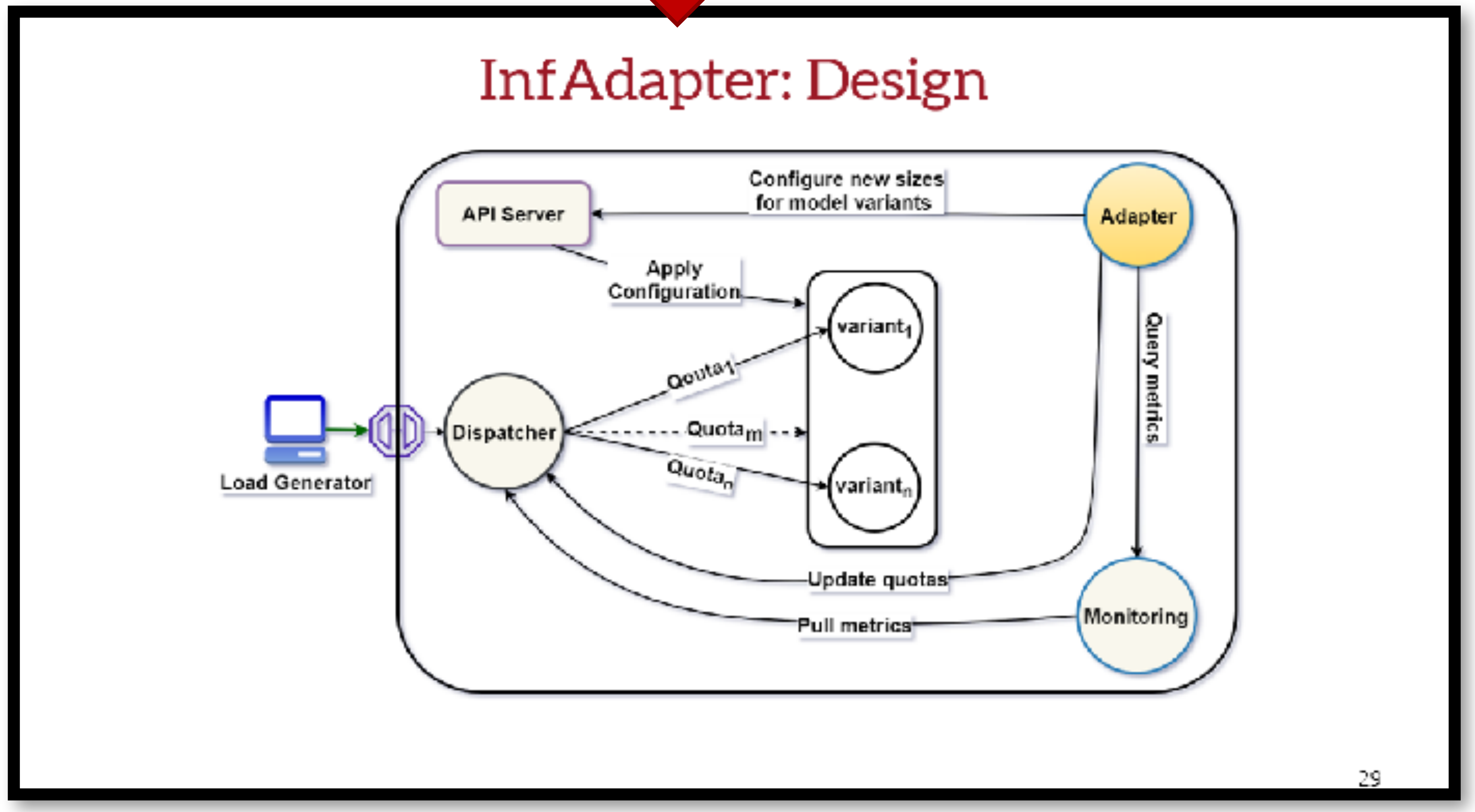
InfAdapter!

https://github.com/reconfigurable-ml-pipeline/InfAdapter

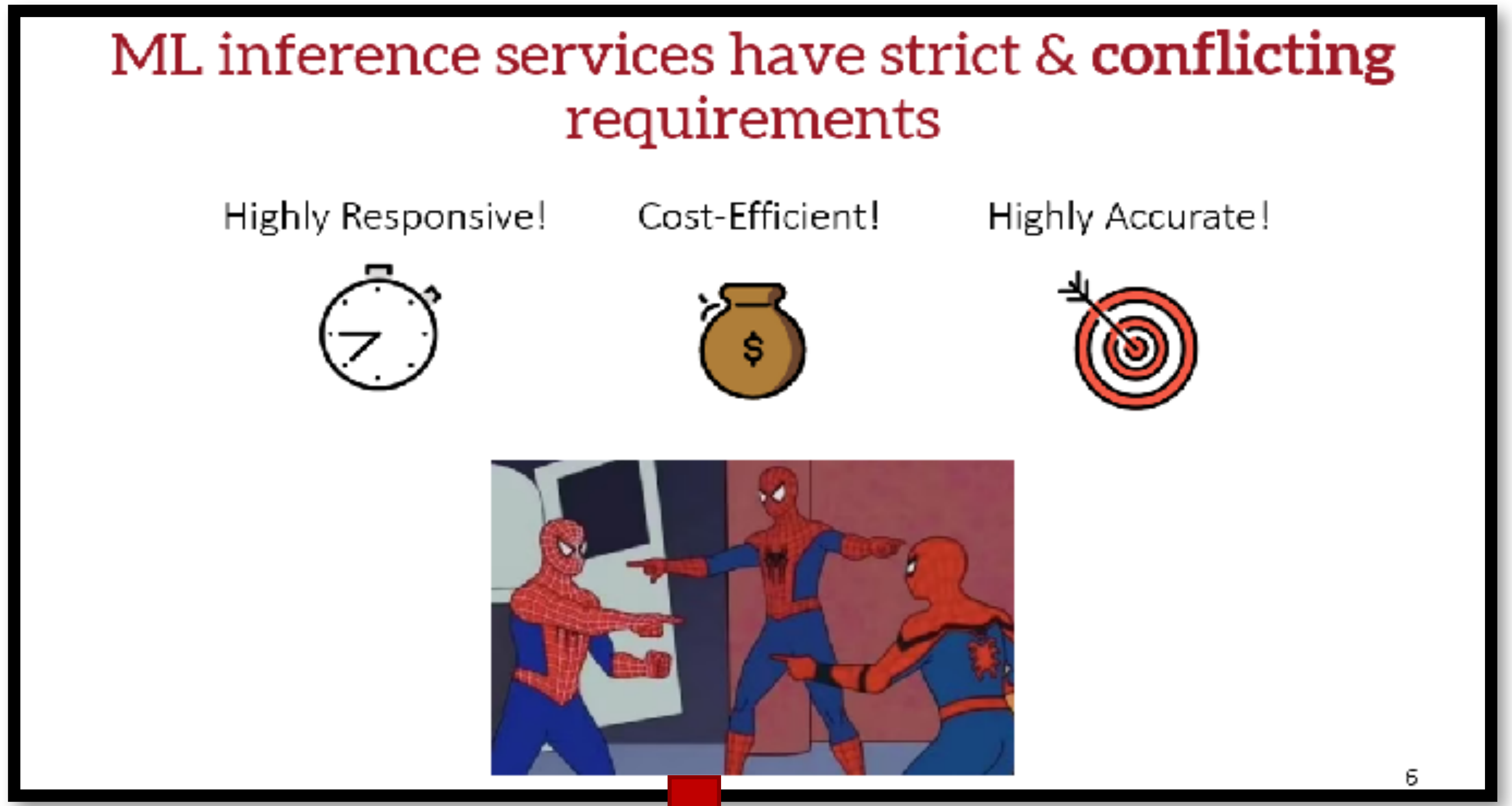


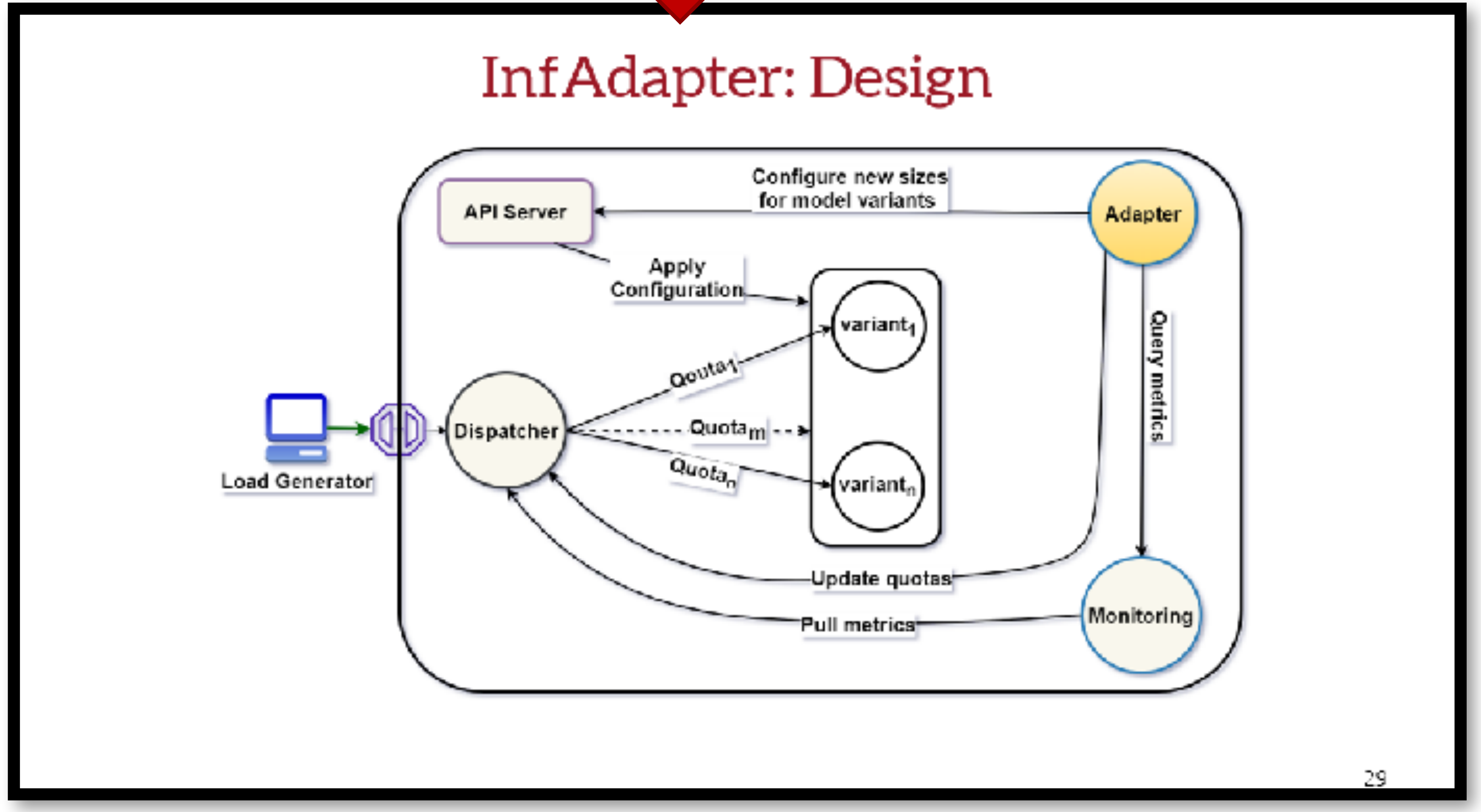
https://github.com/reconfigurable-ml-pipeline/InfAdapter

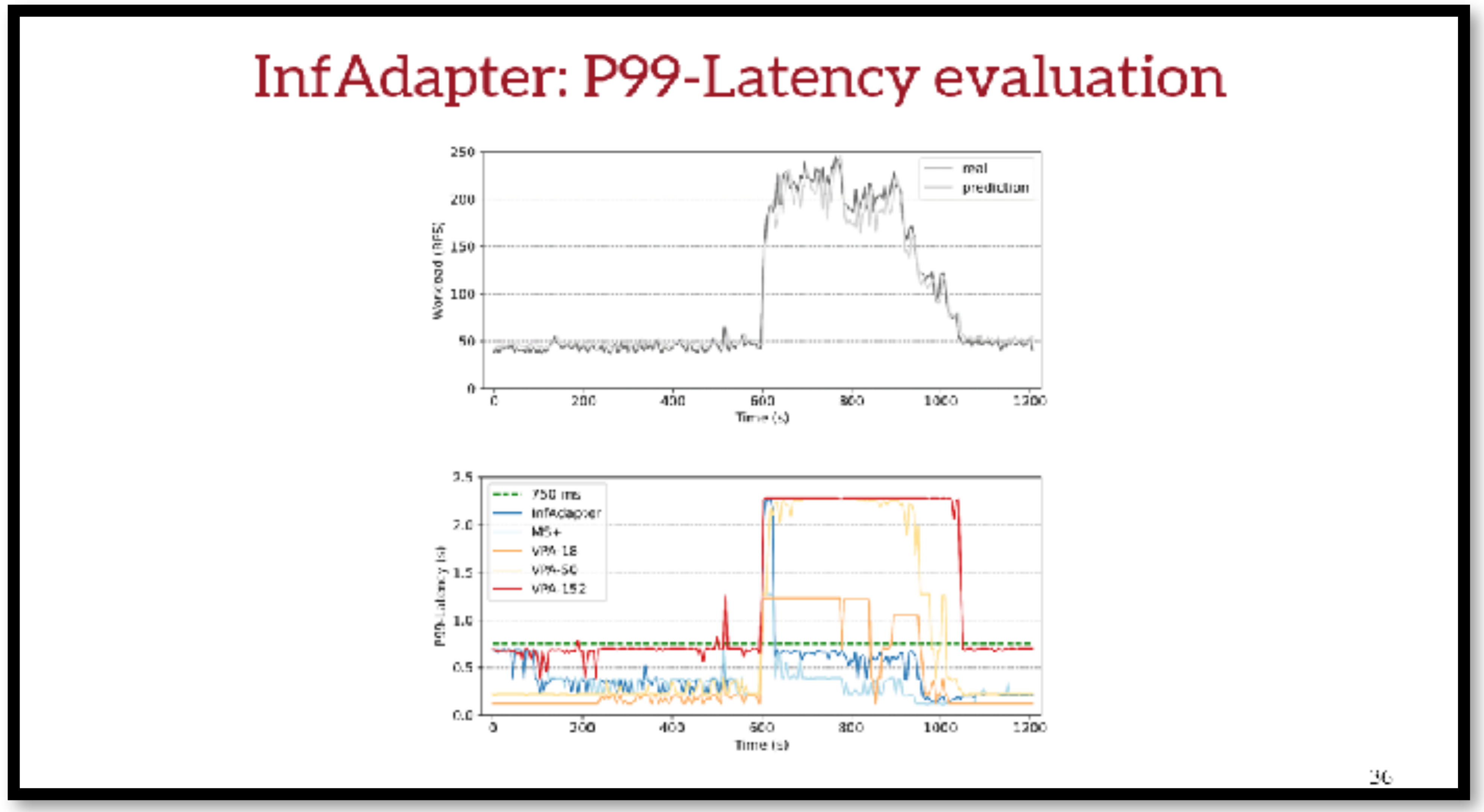




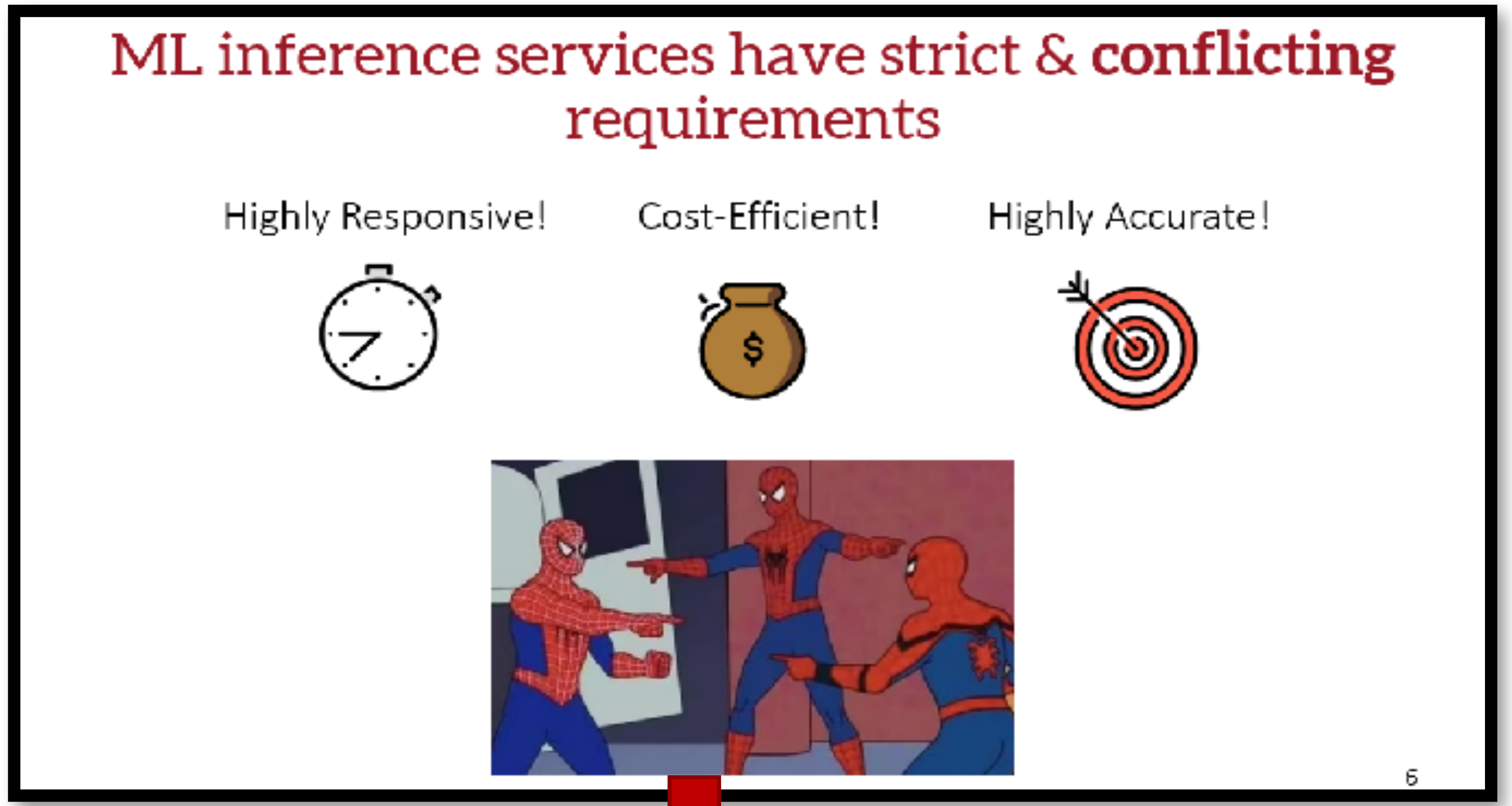
https://github.com/reconfigurable-ml-pipeline/InfAdapter

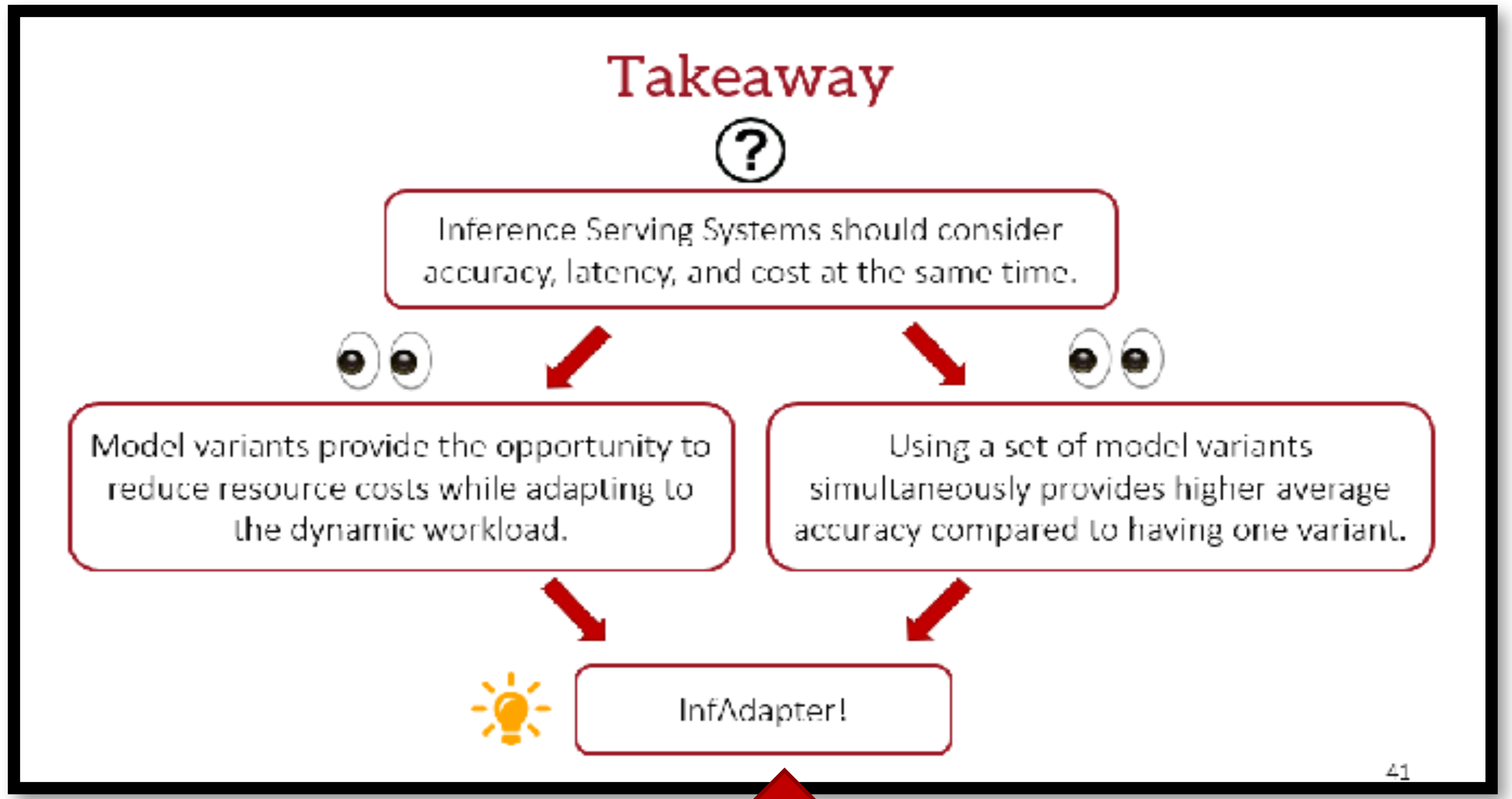


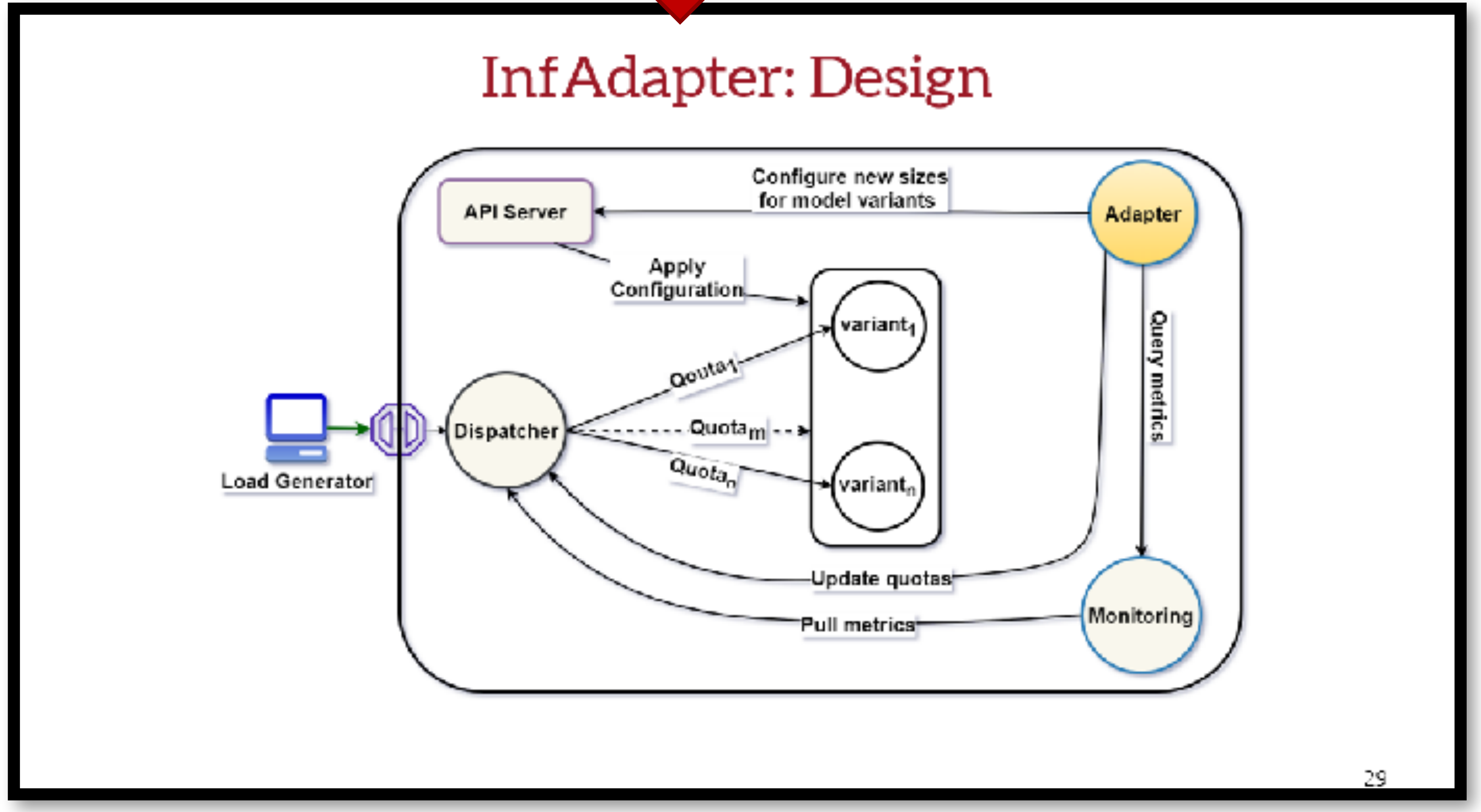


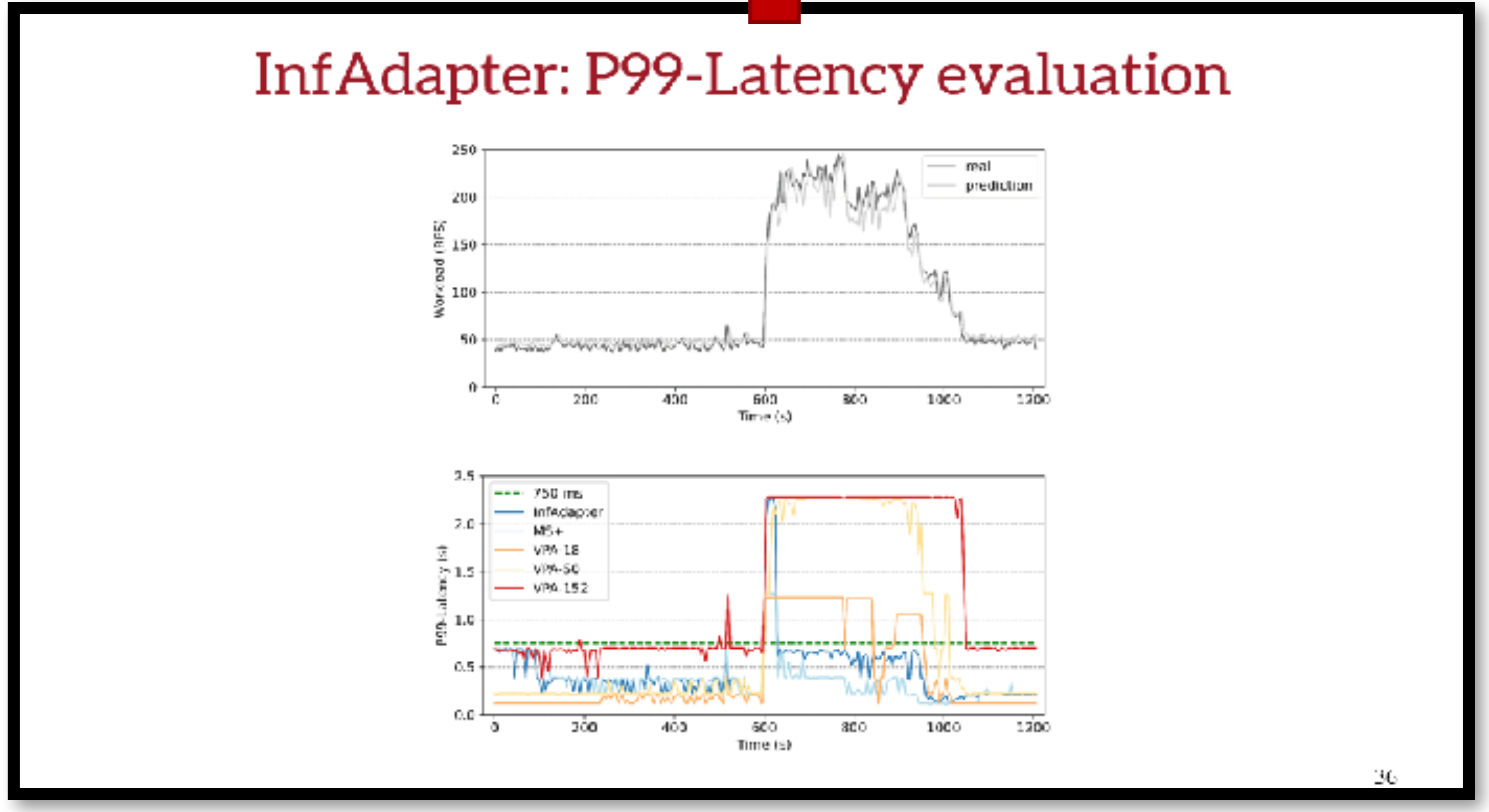


https://github.com/reconfigurable-ml-pipeline/InfAdapter









EuroMLSys

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

Mehran Salmani*, Saeid Ghafouri§‡, Alireza Sanaee§, Kamran Razavi†, Max Mühlhäuser†, Joseph Doyle§, Pooyan Jamshidi‡, Mohsen Sharifi*

Iran University of Science and Technology*, Queen Mary University of London§, Technical University of Darmstadt†, University of South Carolina‡

InfAdapter [2023]: Autoscaling for ML **Model Inference**

JSys Journal of Systems Research Volume 4, Issue 1, April 2024

[SOLUTION] IPA: INFERENCE PIPELINE ADAPTATION TO ACHIEVE HIGH ACCURACY AND COST-EFFICIENCY

Saeid Ghafouri
University of South Carolina & Queen Mary University of London

Kamran Razavi
Technical University of Darmstadt

Mehran Salmani
Technical University of Ilmenau

Alireza Sanaee
Queen Mary University of London

Tania Lorido Botran
Roblox

Lin Wang
Paderborn University

Joseph Doyle
Queen Mary University of London

Pooyan Jamshidi
University of South Carolina

IPA [2024]: Autoscaling for **ML Inference Pipeline**

EuroMLSys

Sponge: Inference Serving with Dynamic SLOs Using In-Place Vertical Scaling

Kamran Razavi*
Technical University of Darmstadt

Saeid Ghafouri*
Queen Mary University of London

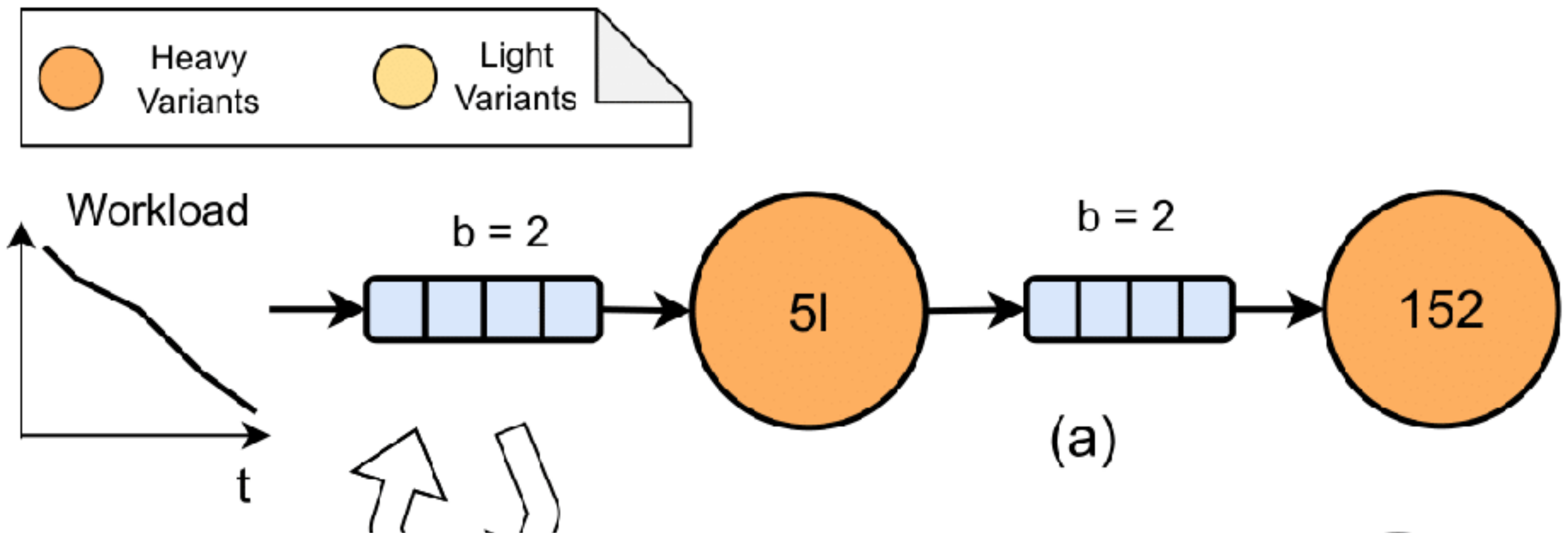
Max Mühlhäuser
Technical University of Darmstadt

Pooyan Jamshidi
University of South Carolina

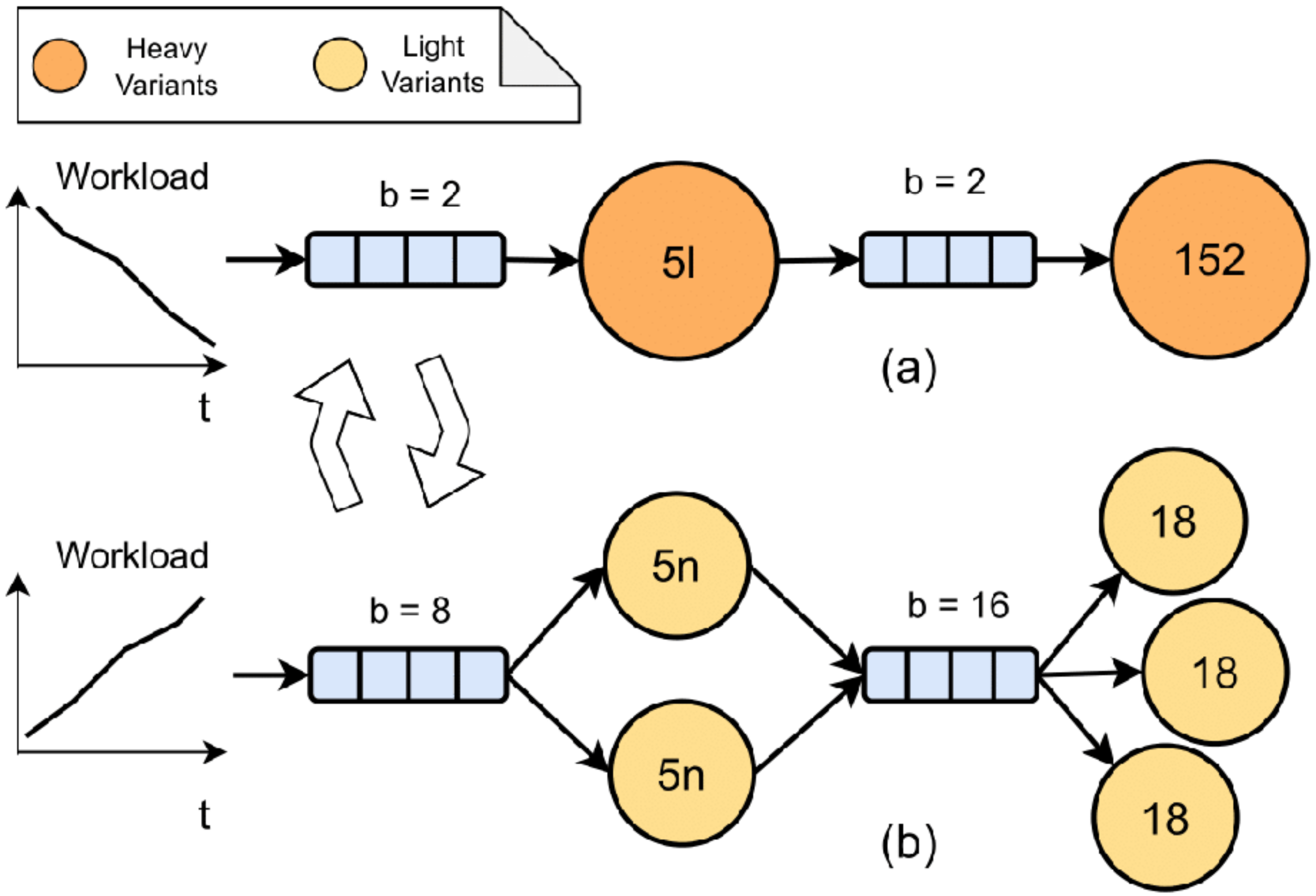
Lin Wang
Paderborn University

Sponge [2024]: Autoscaling for ML Inference Pipeline with **Dynamic SLO**

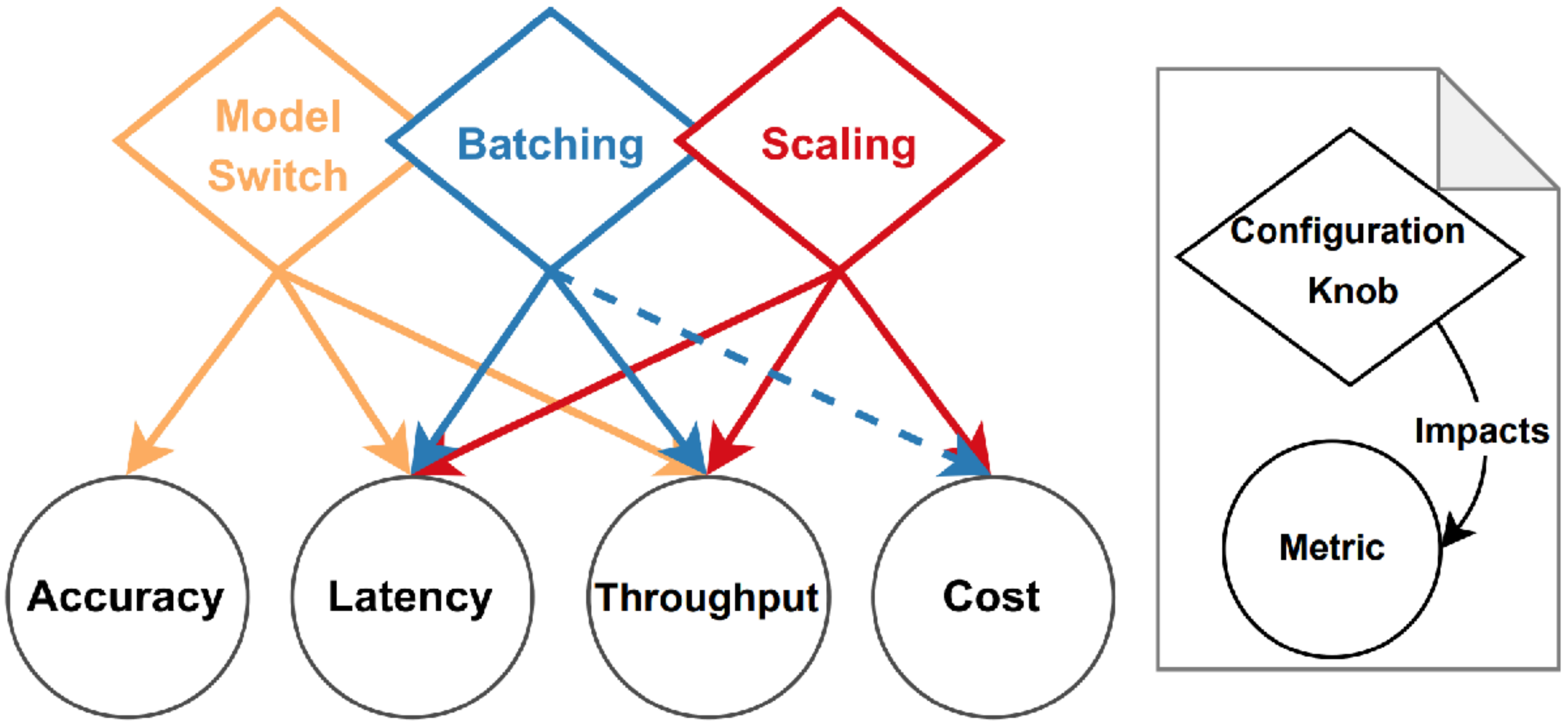
The Variabilities ML Pipelines



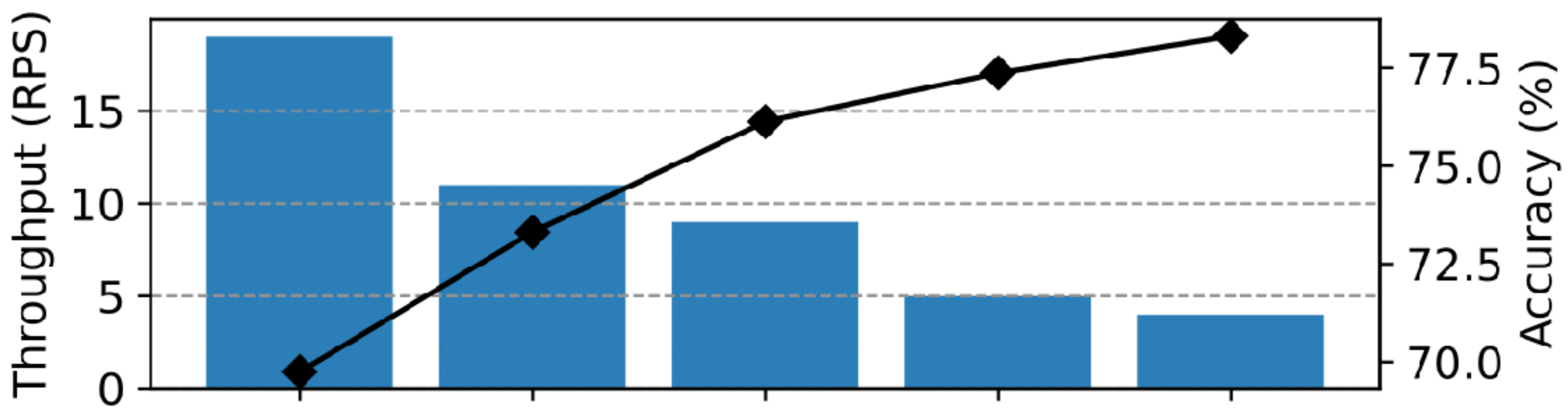
The Variability Space of Multi-Node ML Pipelines is Much Larger than a Single-Node Pipelines



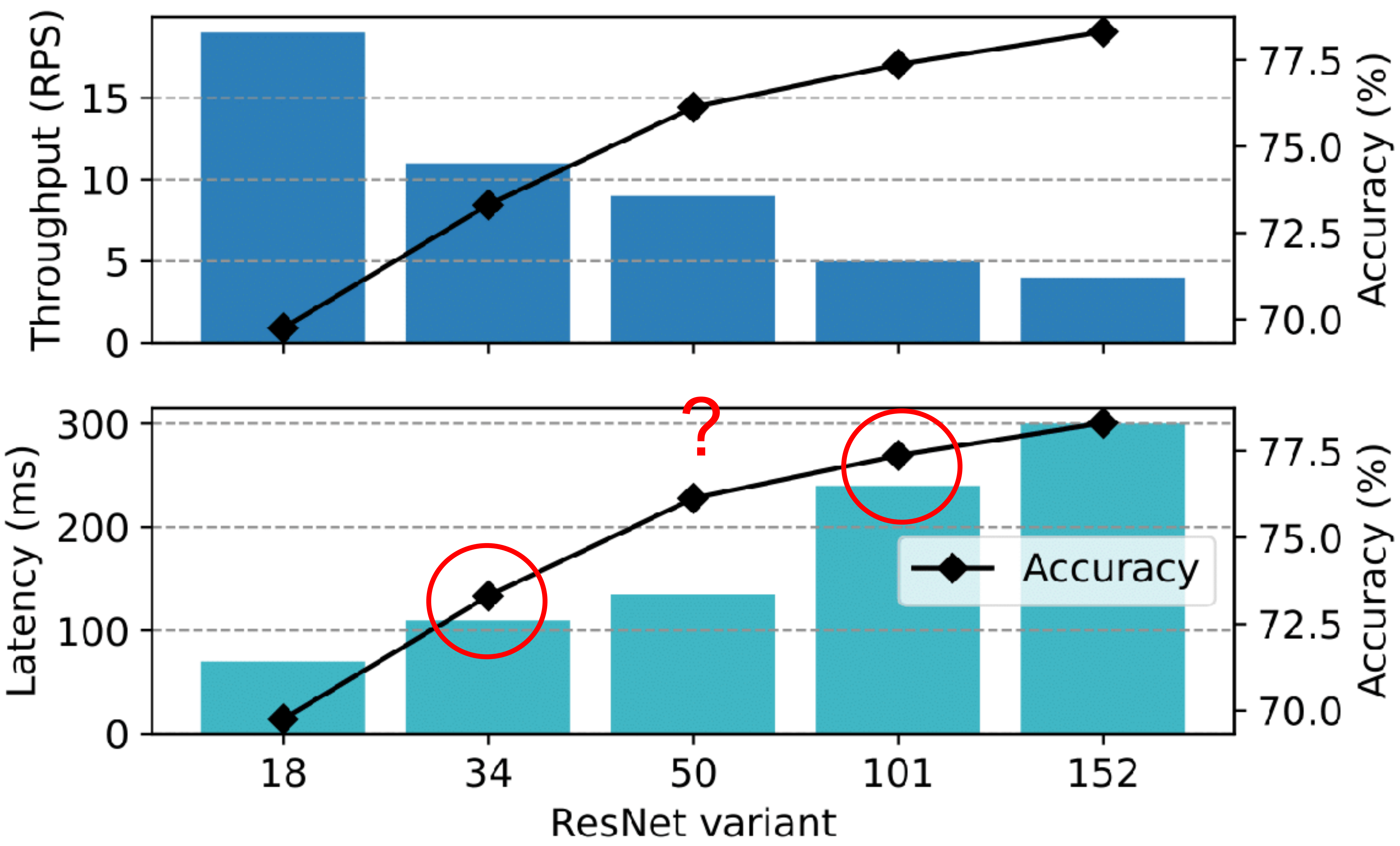
Search Space



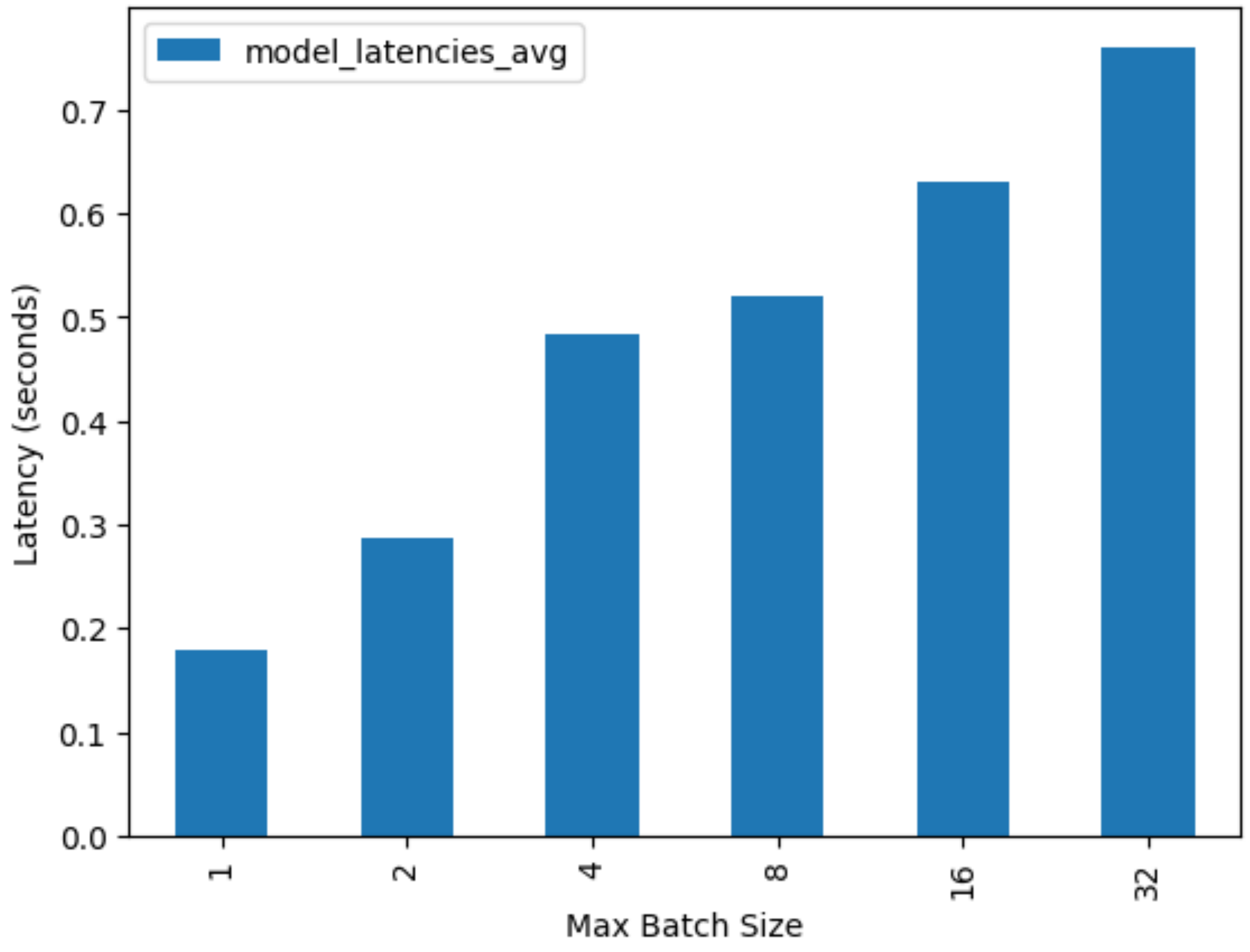
Is only scaling enough?

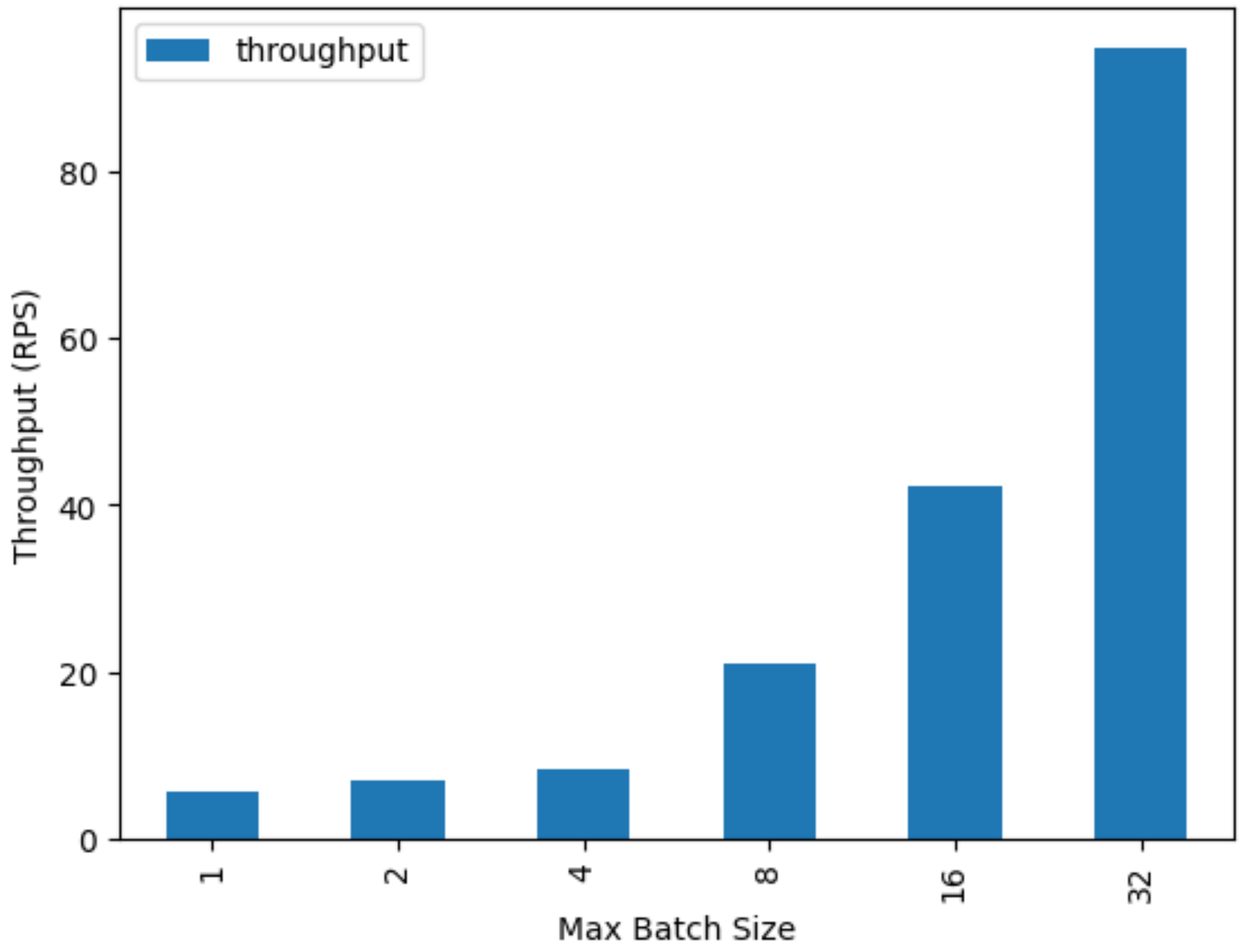


Is only scaling enough?

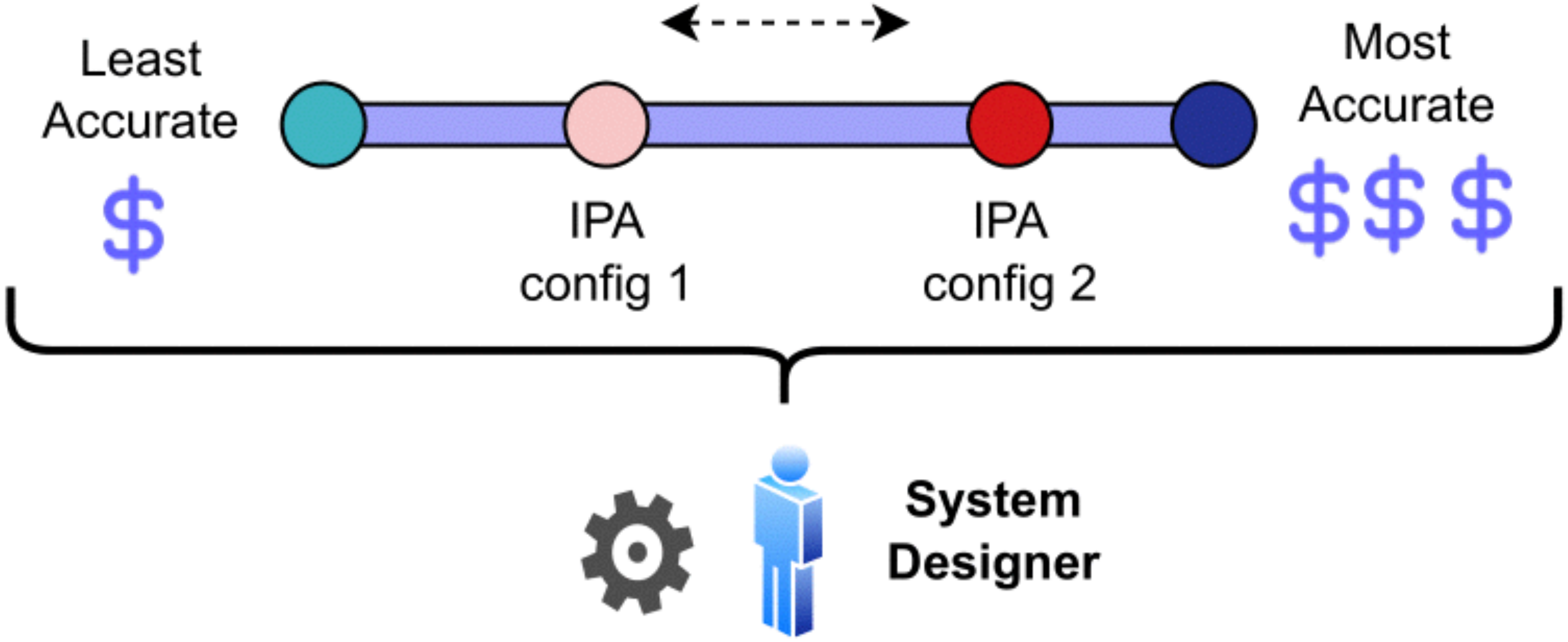


Effect of Batching





Goal: Providing a flexible inference pipeline



Problem Formulation

$$f(n,s,I) = \alpha \sum_{s \in P} \Big(\sum_{m \in M_s} a_{s,m} . I_{s,m} \Big)$$

Accuracy Objective

$$- \beta \sum_{s \in P} n_s . R_s$$

Resource Objective

$$- \delta \sum_{s \in P} b_s$$

Batch Control

Problem Formulation

Latency SLA

$$
\begin{aligned}
\max \quad & f(n,s,I) \\
\text{subject to} \quad & \sum_{s\in P} l_s(b_s) + q_s(b_s) \le SLA_P, \\
& \text{if } I_{s,m} = 1, \text{ then} \\
& \quad n_s \cdot h_s(b_s) \ge \lambda_p, \quad \forall s \in P \\
& \sum_{m\in M_s} I_{s,m} = 1, \quad \forall s \in P \\
& n_s, b_s \in \mathbb{Z}^+, \quad I_{s,m} \in \{0,1\}, \quad \forall s \in S
\end{aligned}
$$

Throughput Constraint

One active model per node

$$
\begin{aligned}
f(n,s,I) = \alpha & \sum_{s\in P} \Big(\sum_{m\in M_s} a_{s,m}.I_{s,m} \Big) \\
& - \beta \sum_{s\in P} n_s.R_s \\
& - \delta \sum_{s\in P} b_s
\end{aligned}
$$

Evaluations

Setup and Partial Results

For more comprehensive results, please refer to the IPA paper!

How to navigate Model Variants

kubernetes

1. Industry standard
2. Used in recent research
3. Complete set of autoscaling, scheduling, observability tools (e.g. CPU usage)
4. APIs for changing the current AutoScaling algorithms

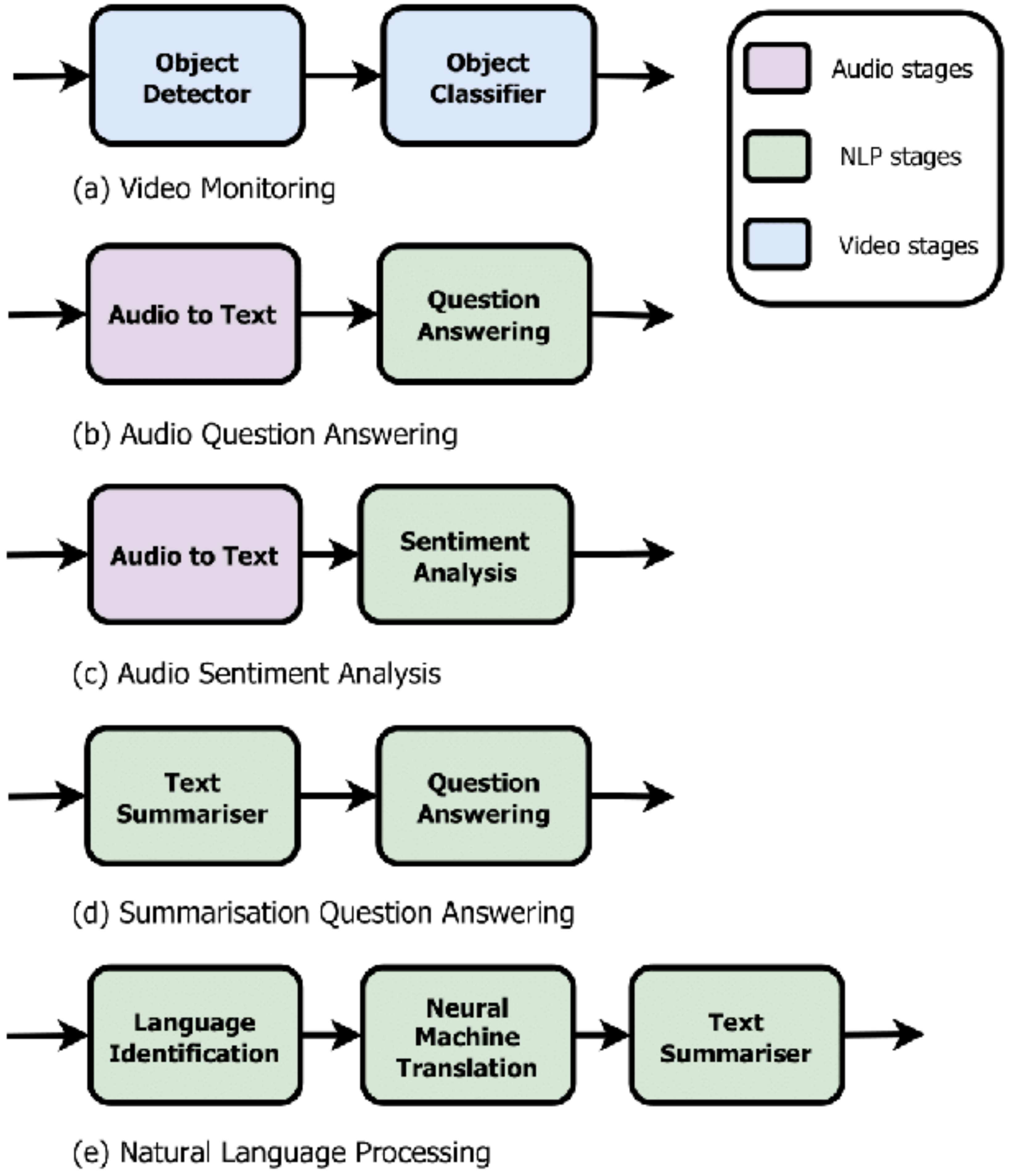
CORE

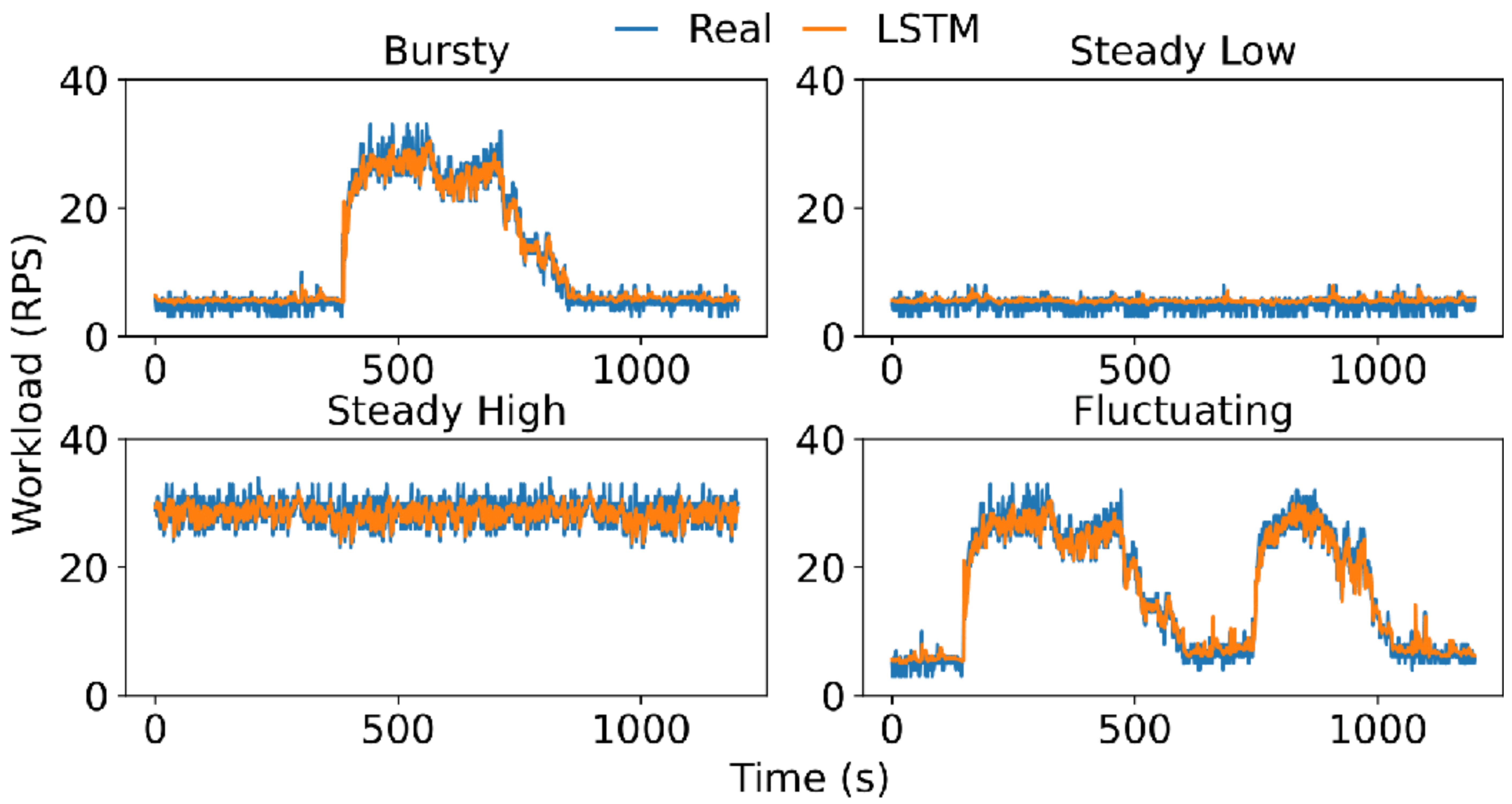
1. Industry standard ML server
2. Have the ability make inference graph
3. Rest and GRPC endpoints
4. Have many of the features we need like monitoring stack out of the box

Evaluation

https://github.com/reconfigurable-ml-pipeline/ipa

kubernetes CORE





We compared IPA with RIM and FA2

Rim: Offloading Inference to the Edge

Yitao Hu
University of Southern California
yitaoh@usc.edu

Weiwu Pang
University of Southern California
weiwupan@usc.edu

Xiaochen Liu
University of Southern California
liu851@usc.edu

Rajrup Ghosh
University of Southern California
rajrupgh@usc.edu

Bongjun Ko
IBM Research
bongjun_ko@us.ibm.com

Wei-Han Lee
IBM Research
wei-han.lee1@ibm.com

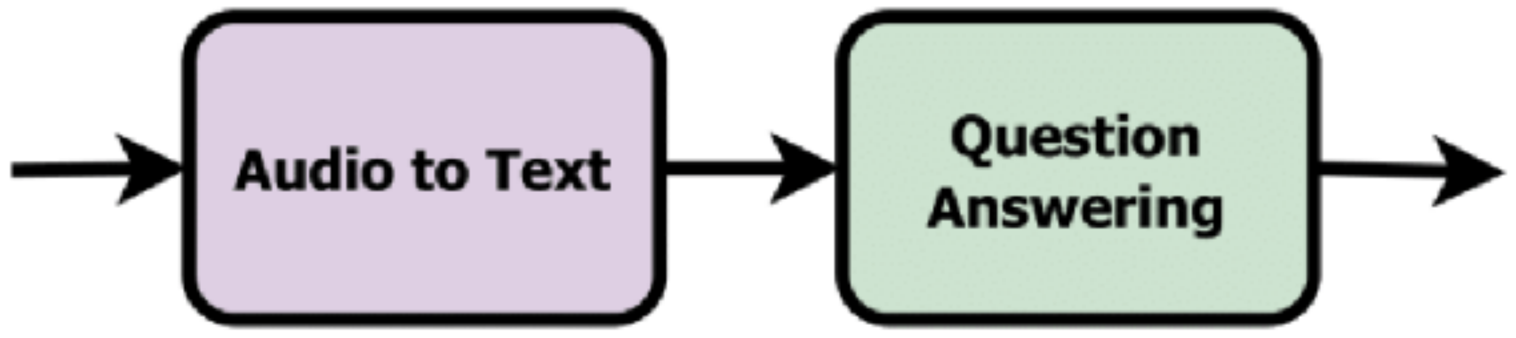
Ramesh Govindan
University of Southern California
ramesh@usc.edu

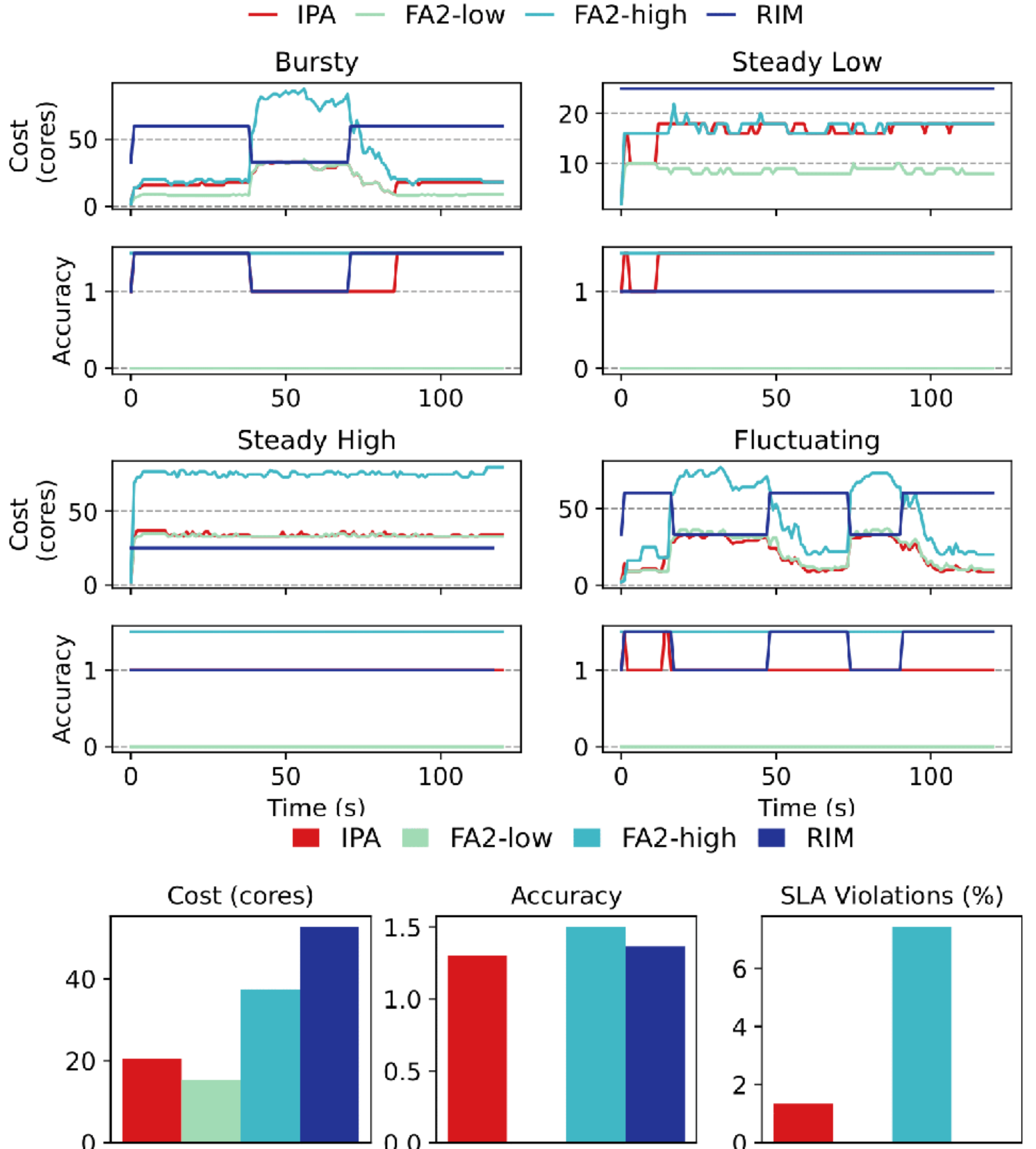
FA2: Fast, Accurate Autoscaling for Serving Deep Learning Inference with SLA Guarantees

Kamran Razavi[†], Manisha Luthra[†], Boris Koldehofe[†,‡], Max Mühlhäuser[†], Lin Wang[†,§]

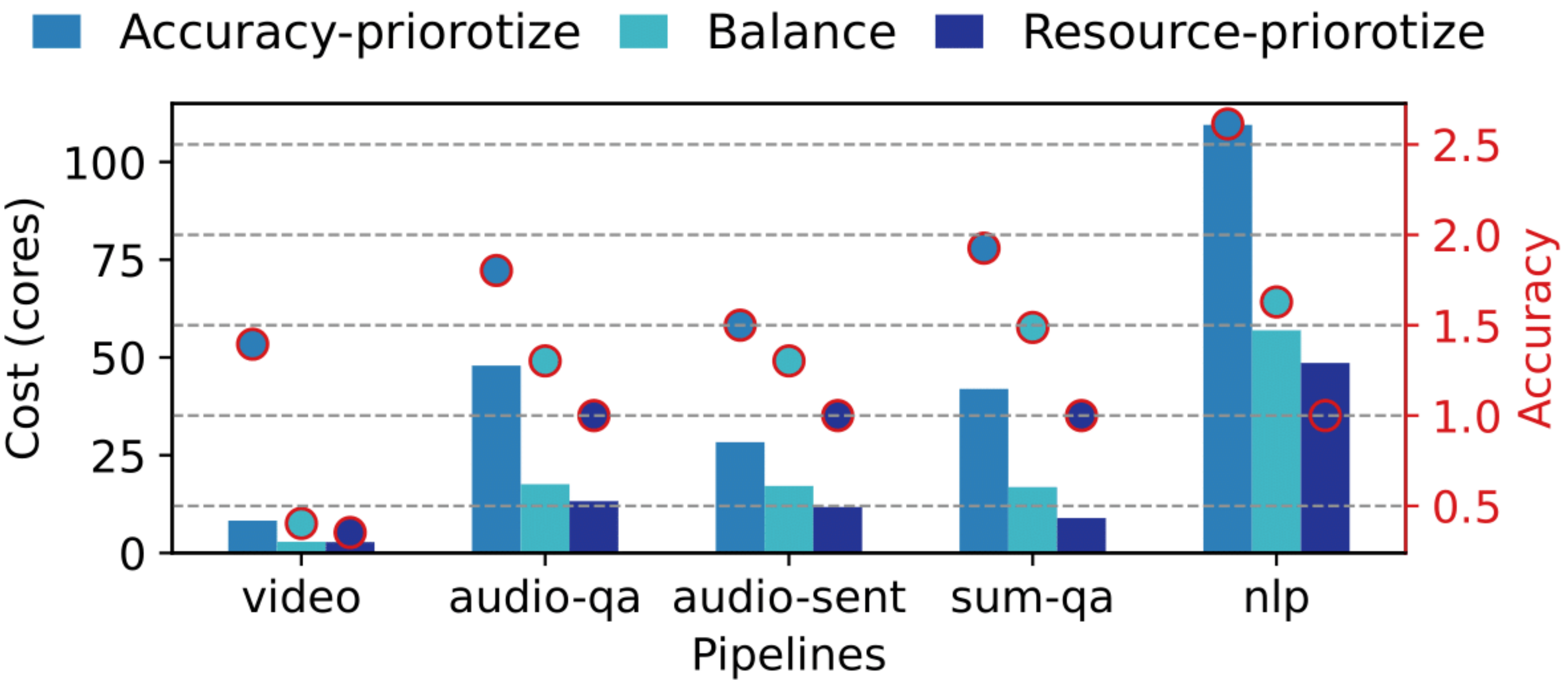
[†]Technische Universität Darmstadt
[‡]University of Groningen
[§]Vrije Universiteit Amsterdam

Audio + QA Pipeline

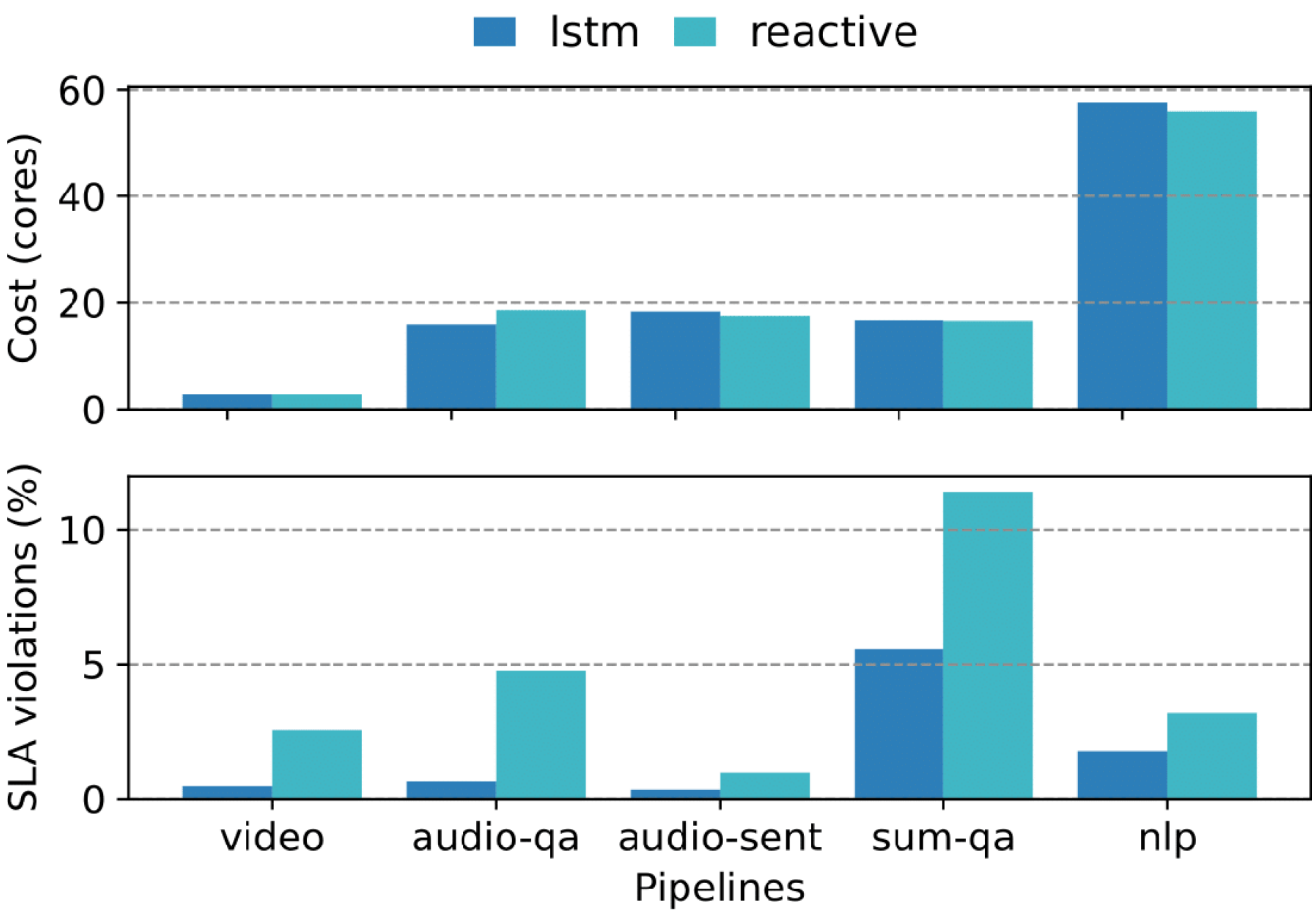




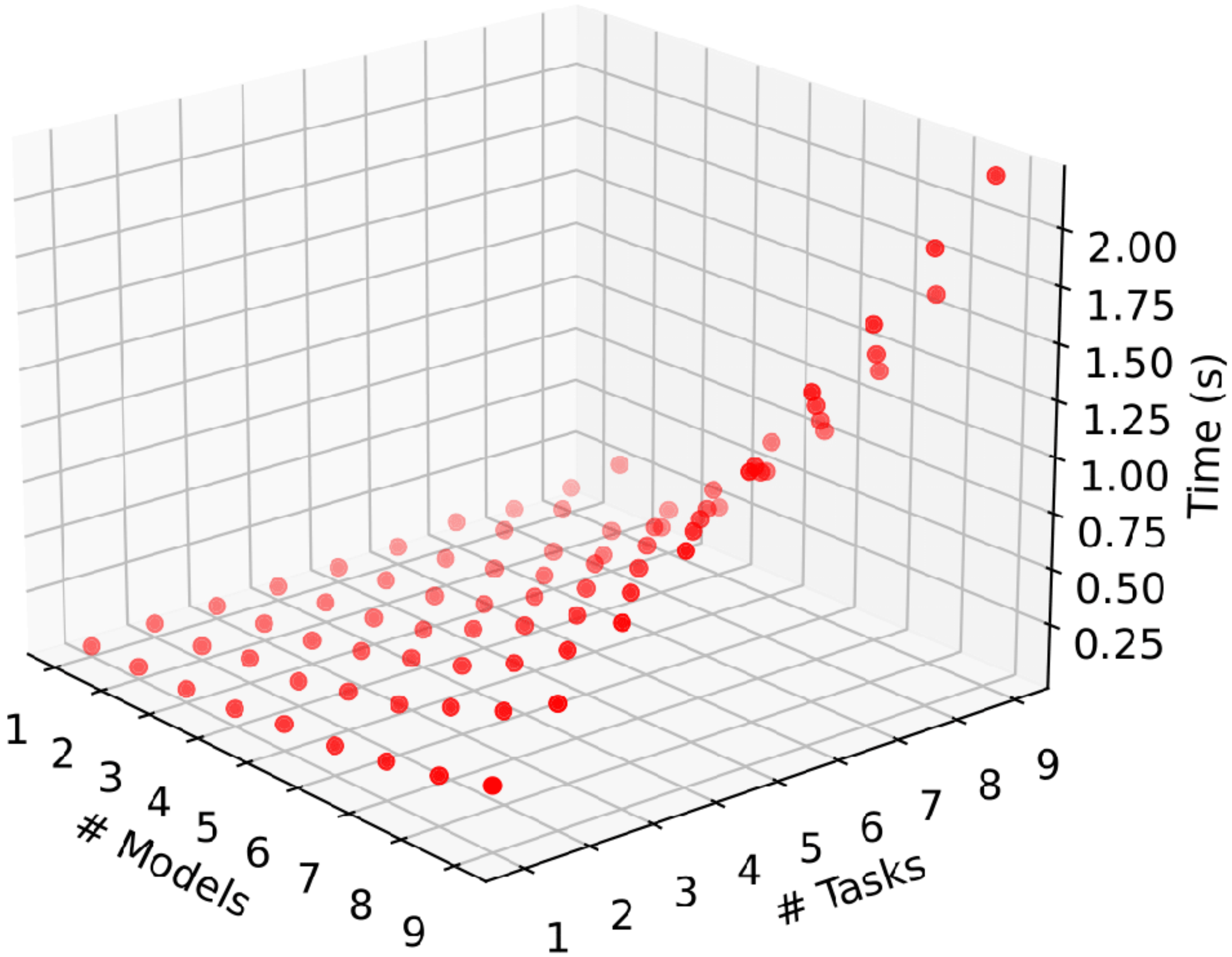
Adaptivity to multiple objectives



Effect of predictor



Gurobi solver scalability



Full replication package is available

https://github.com/reconfigurable-ml-pipeline

AdaptiveFlow

Repositories related to Sustainability, Performance, Auto-scaling, Reconfiguration, Runtime Optimizations for ML Inference Pipelines

1 follower · United States of America

Unfollow

Popular repositories

ipa — Public
Source code of IPA
Jupyter Notebook · 8 · 4

InfAdapter — Public
Source code of "Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems"
Python · 7

load_tester — Public
Python · 2

kubernetes-python-client — Public
Python

INFaaS — Public
Forked from stanford-mast/INFaaS
Model-less Inference Serving
C++

View as: Public

You are viewing the README and pinned repositories as a public user.

You can create a README file or pin repositories visible to anyone.

Get started with tasks that most successful organizations complete.

Discussions

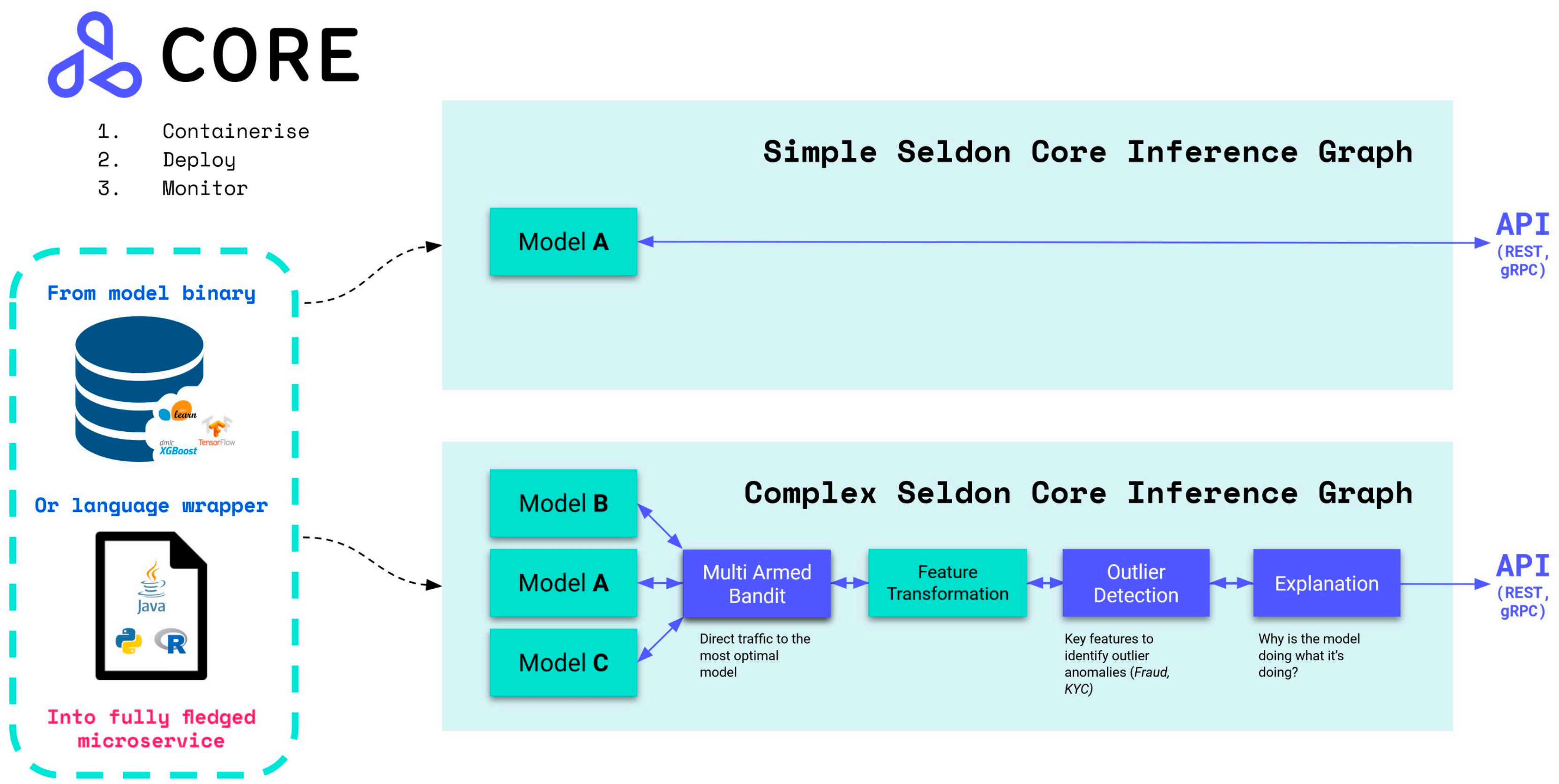
Set up discussions to engage with your community!

Turn on discussions

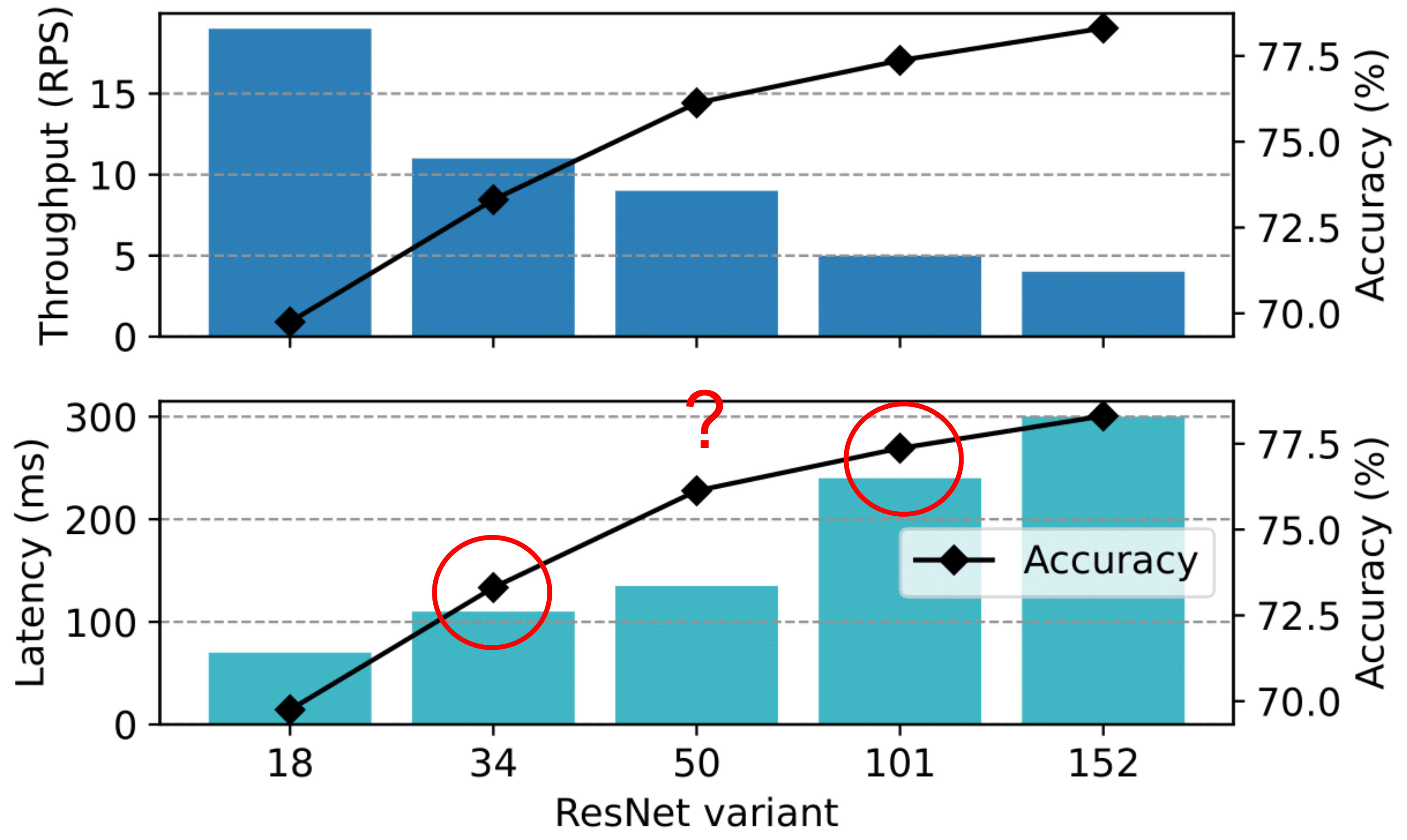
People

Model Serving

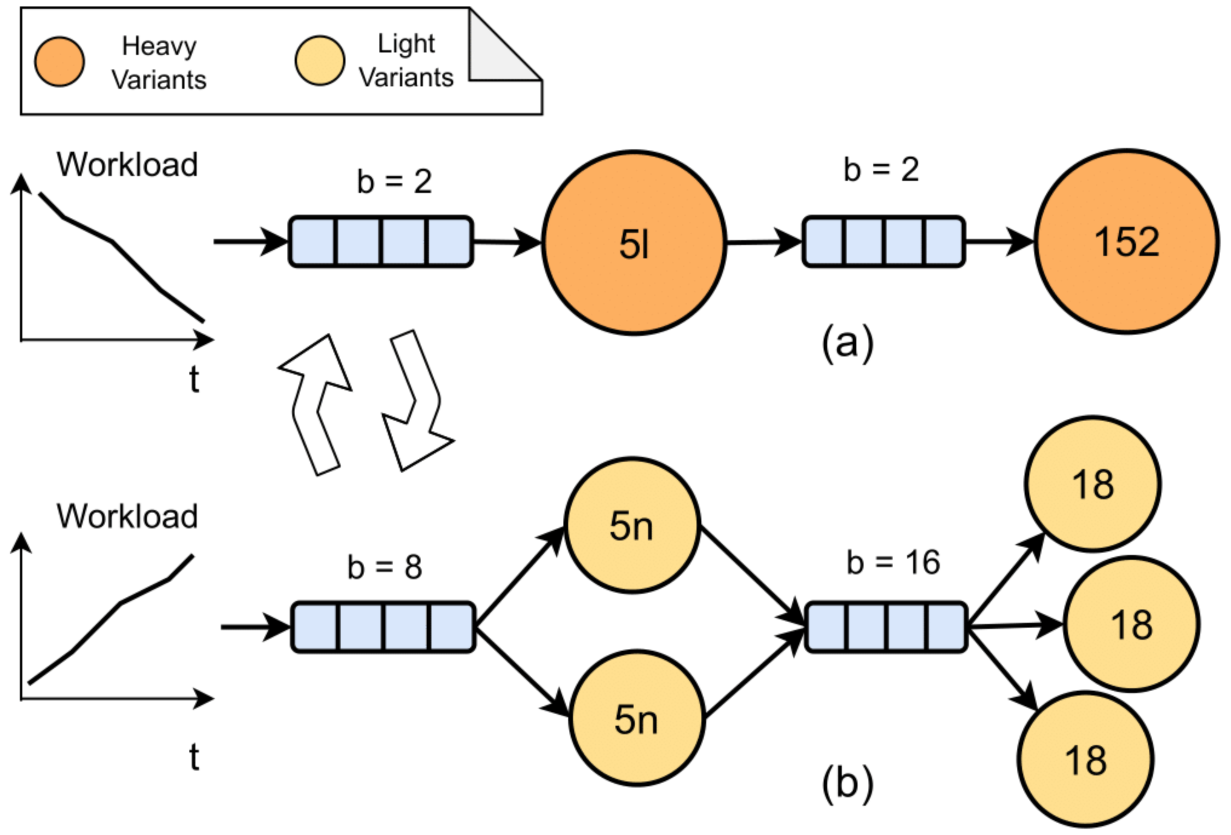
Pipeline



Is only scaling enough?

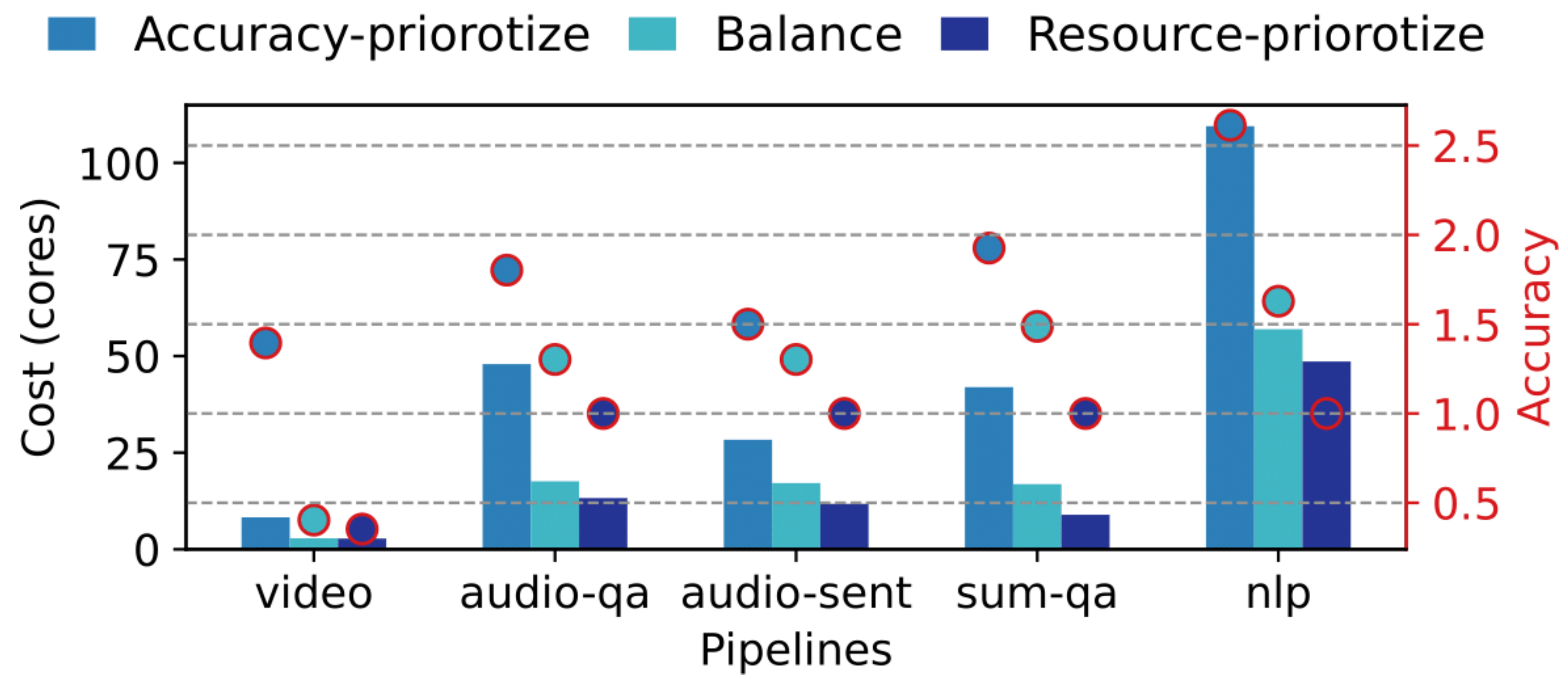


Snapshot of the System



X

Adaptivity to multiple objectives



X

EuroMLSys

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

Mehran Salmani[*], Saeid Ghafouri[§‡], Alireza Sanaee[§], Kamran Razavi[†], Max Mühlhäuser[†], Joseph Doyle[§], Pooyan Jamshidi[‡], Mohsen Sharifi[*]

Iran University of Science and Technology[*], Queen Mary University of London[§], Technical University of Darmstadt[†], University of South Carolina[‡]

InfAdapter [2023]: Autoscaling for ML **Model Inference**

JSys Journal of Systems Research Volume 4, Issue 1, April 2024

[SOLUTION] IPA: INFERENCE PIPELINE ADAPTATION TO ACHIEVE HIGH ACCURACY AND COST-EFFICIENCY

Saeid Ghafouri
University of South Carolina & Queen Mary University of London

Kamran Razavi
Technical University of Darmstadt

Mehran Salmani
Technical University of Ilmenau

Alireza Sanaee
Queen Mary University of London

Tania Lorido Botran
Roblox

Lin Wang
Paderborn University

Joseph Doyle
Queen Mary University of London

Pooyan Jamshidi
University of South Carolina

IPA [2024]: Autoscaling for **ML Inference Pipeline**

EuroMLSys

Sponge: Inference Serving with Dynamic SLOs Using In-Place Vertical Scaling

Kamran Razavi[*]
Technical University of Darmstadt

Saeid Ghafouri[*]
Queen Mary University of London

Max Mühlhäuser
Technical University of Darmstadt

Pooyan Jamshidi
University of South Carolina

Lin Wang
Paderborn University

Sponge [2024]:
Autoscaling for
ML Inference Pipeline with
Dynamic SLO

Dynamic User -> Dynamic Network Bandwidths

- Users move
  - Fluctuations in the network bandwidths
    - Reduced time-budget for processing requests

SLO

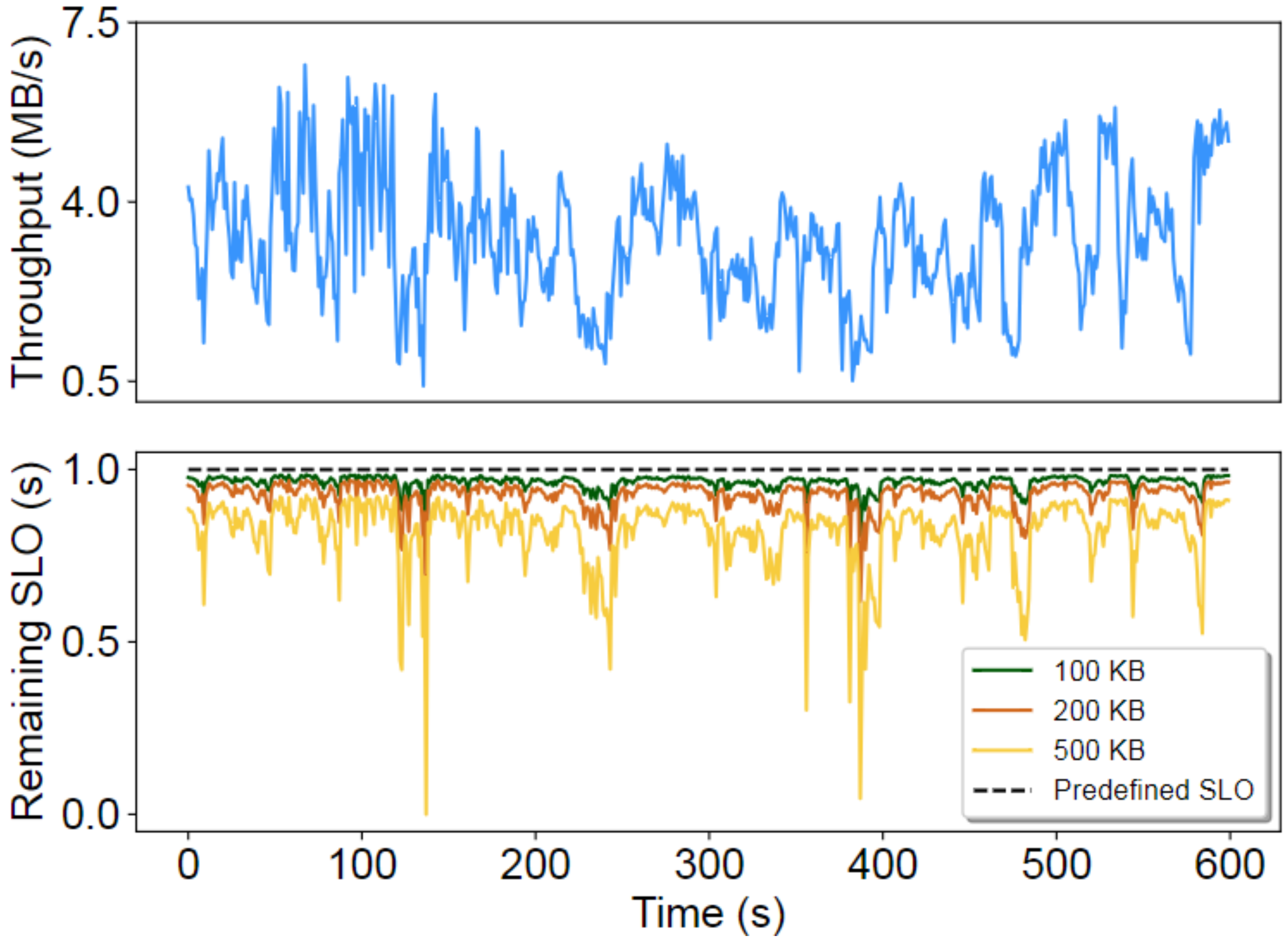
network latency processing latency

Dynamic User -> Dynamic Network Bandwidths

- Users move
  - Fluctuations in the network bandwidths
    - Reduced time-budget for processing requests

SLO

network latency processing latency



Inference Serving Requirements

Highly Responsive!
(end-to-end latency guarantee)

Cost-Efficient!
(least resource consumption)

Resource Scaling

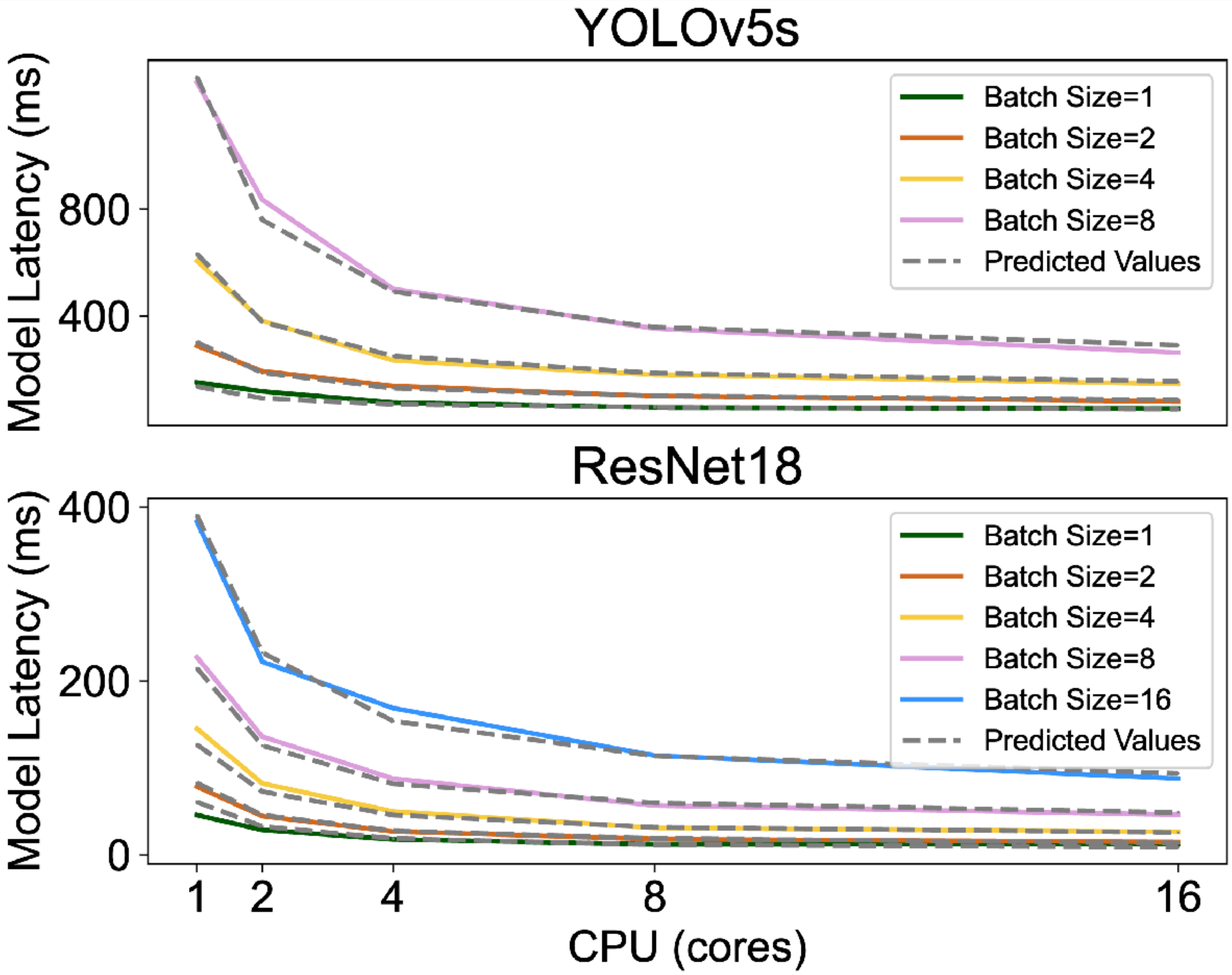
Sponge!

In-place Vertical Scaling
(more responsive)

Horizontal Scaling
(more cost efficient)

Vertical Scaling DL Model Profiling

- How much resource should be allocated to a DL model?
  - Latency/batch size → linear relationship
  - Latency/CPU allocation → inverse relationship



Problem Formulation

$$
\begin{aligned}
\text{Minimize} \quad & c + \delta \times b \\
\text{subject to} \quad & l(b, c) + q_r(b, c) + \mathrm{cl}_{max} \leq SLO, \quad \forall r \in R \\
& h(b, c) \geq \lambda \\
& b, c \in \mathbb{Z}^+
\end{aligned}
$$

Problem Formulation

Minimize resource costs

$$
\begin{aligned}
\text{Minimize} \quad & c + \delta \times b \\
\text{subject to} \quad & l(b,c) + q_r(b,c) + \mathrm{cl}_{max} \leq SLO, \quad \forall r \in R \\
& h(b,c) \geq \lambda \\
& b, c \in \mathbb{Z}^+
\end{aligned}
$$

Problem Formulation

Minimize resource costs

$$\text{Minimize} \quad c + \delta \times b$$

Limit the batch size to grow infinitely!

$$\text{subject to} \quad l(b,c) + q_r(b,c) + \mathrm{cl}_{max} \leq SLO, \quad \forall r \in R$$

$$h(b,c) \geq \lambda$$

$$b, c \in \mathbb{Z}^+$$

Problem Formulation

Minimize resource costs

$$\text{Minimize} \quad c + \delta \times b$$

Limit the batch size to grow infinitely!

$$\text{subject to} \quad l(b,c) + q_r(b,c) + \mathrm{cl}_{max} \leq SLO, \quad \forall r \in R$$

$$h(b,c) \geq \lambda$$

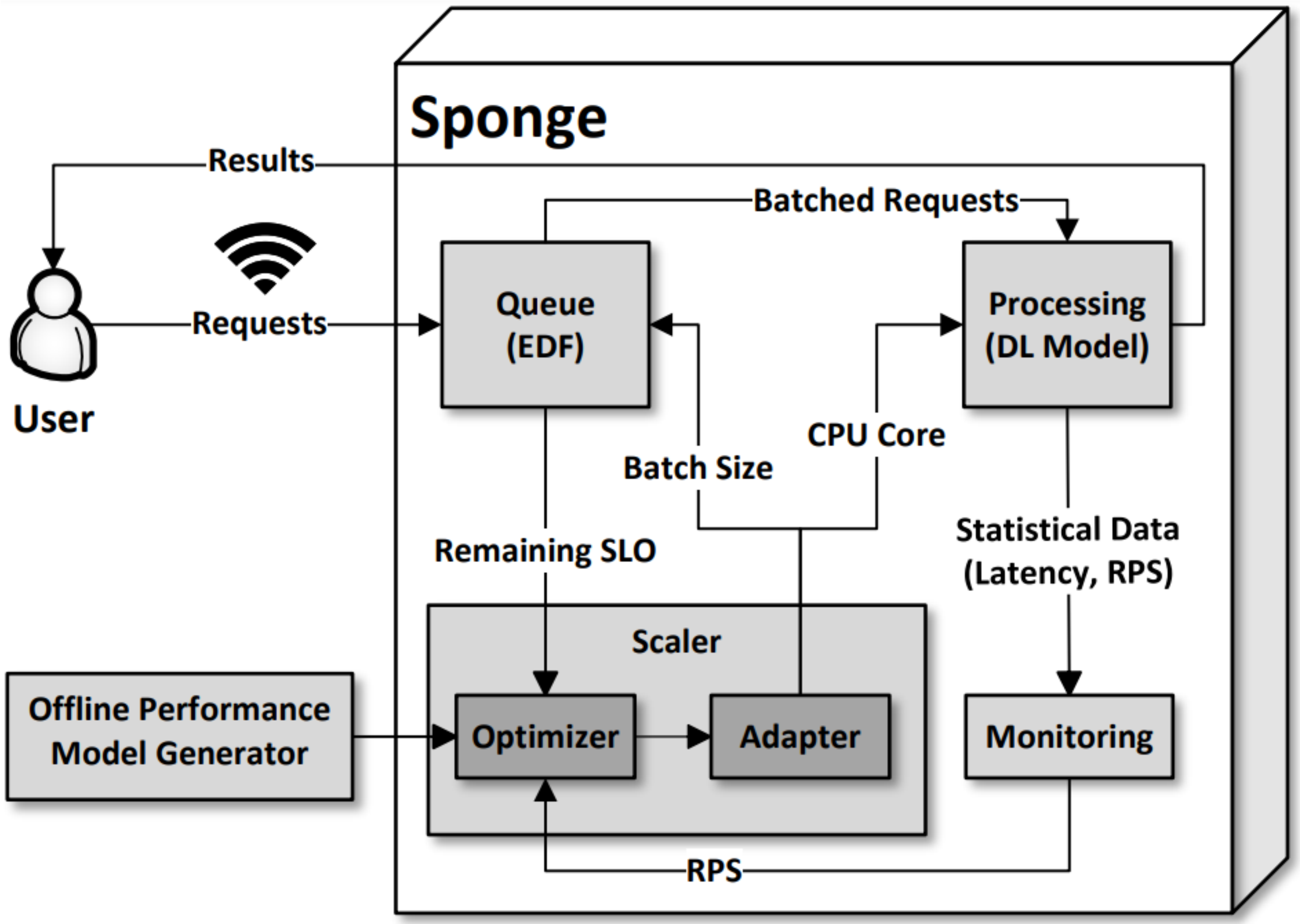
$$b, c \in \mathbb{Z}^+$$

R	Set of all requests
b	Model's batch size
c	Model's CPU allocation
cl_r	Communication latency associated with $r \in R$
cl_{max}	Highest cl_r in R
SLO	Pre-defined SLO for R
$l(b,c)$	Processing time of a model with allocation core c and batch size b
$q_r(b,c)$	Queuing time of $r \in R$ with allocation core c and batch size b
$h(b,c)$	Throughput of a model with allocation core c and batch size b
λ	Request arrival rate

System Design

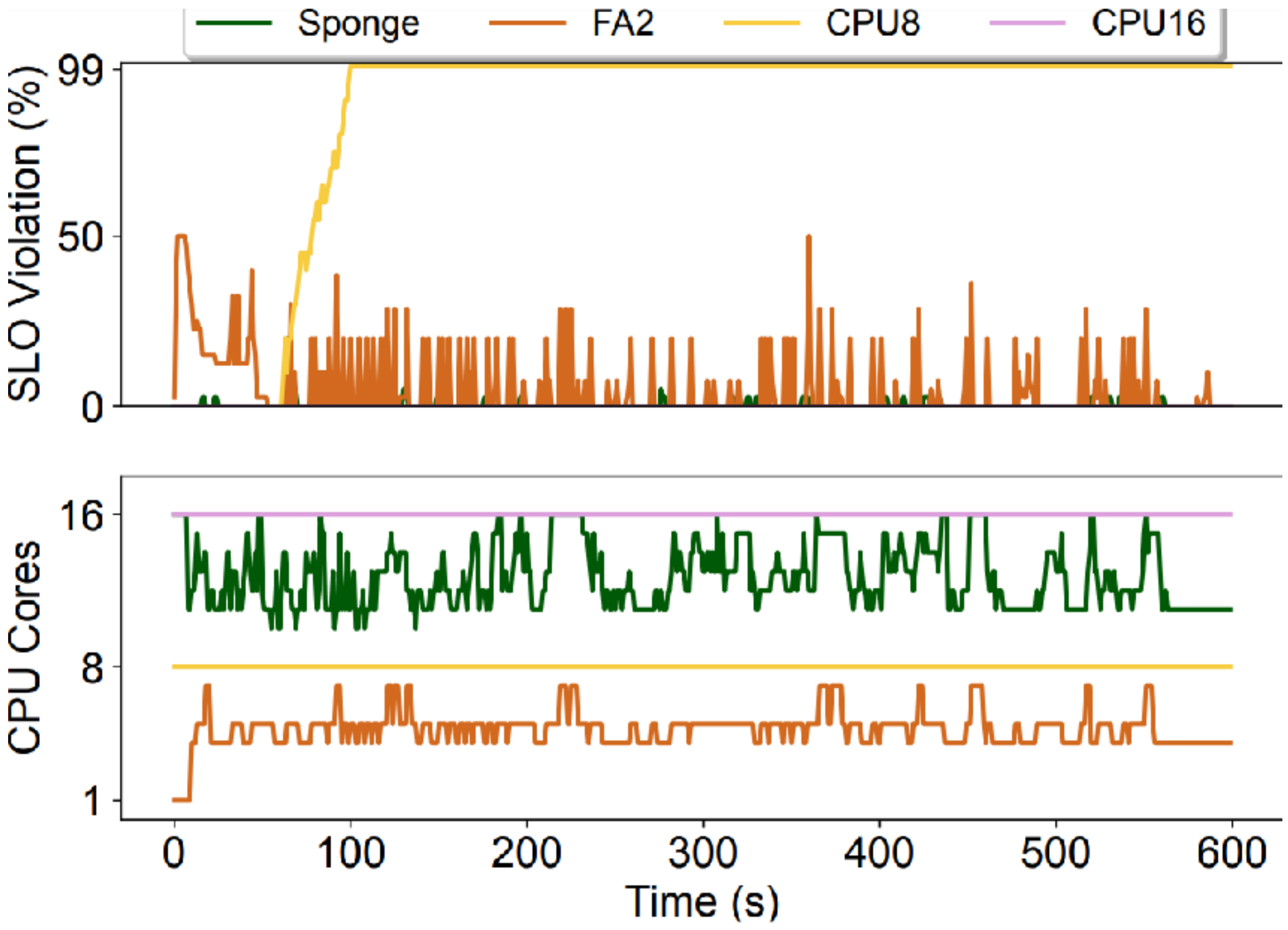
3 design choices:

1. **In-place vertical scaling**
   - Fast response time
2. **Request reordering**
   - High priority requests
3. **Dynamic batching**
   - Increase system utilization



Evaluation

SLO guarantees (99th percentile) with up to 20% resource save up compared to static resource allocation.



Sponge source code:

https://github.com/saeid93/sponge

Future Directions

Resource Scaling

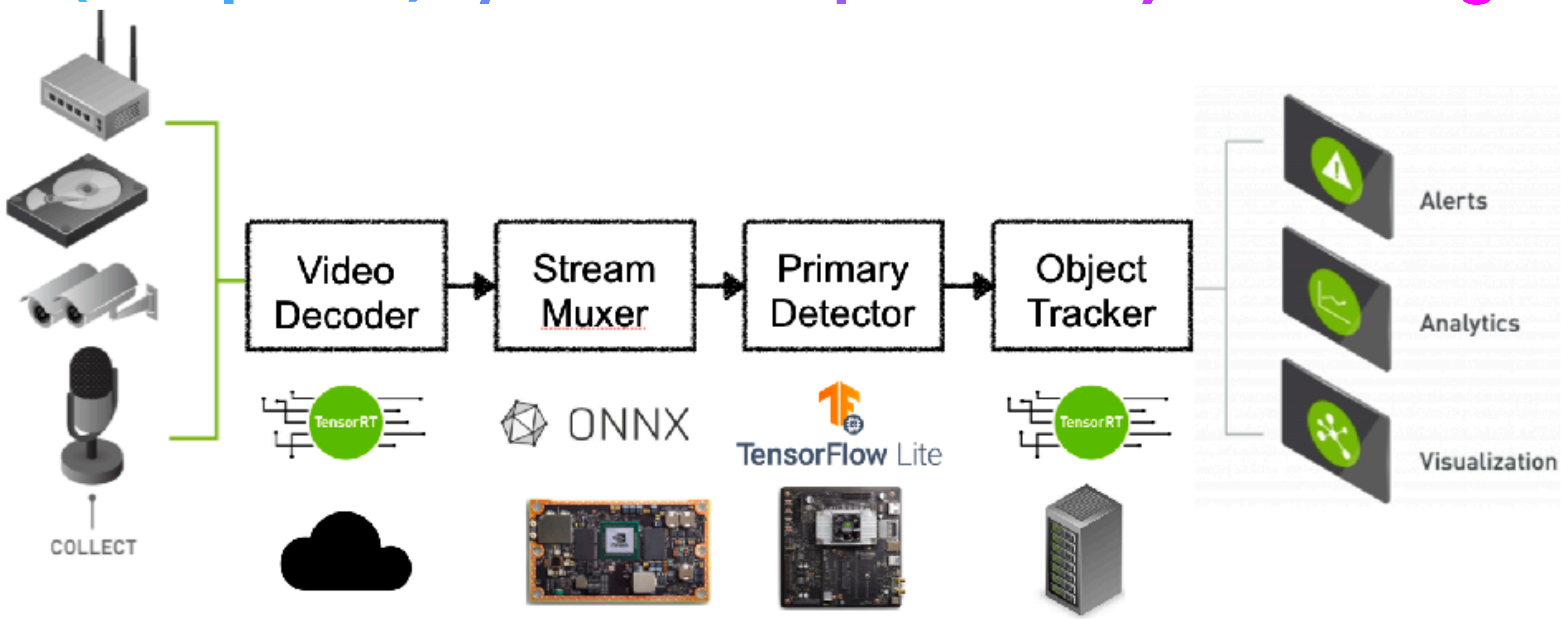
Sponge!

In-place Vertical Scaling
(more responsive)

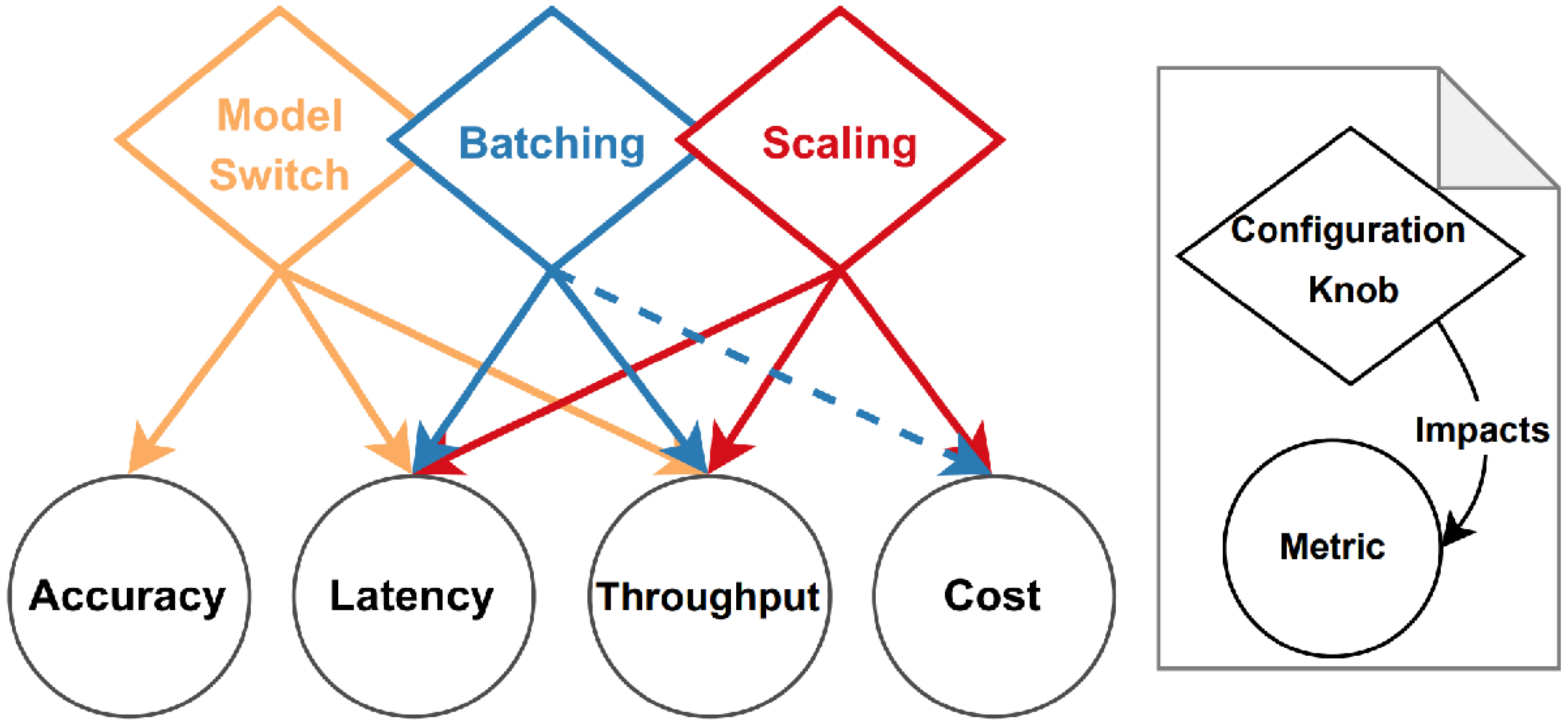
Horizontal Scaling
(more cost efficient)

How can both scaling mechanisms be used jointly under a dynamic workload to be responsive and cost efficient while guaranteeing SLOs?

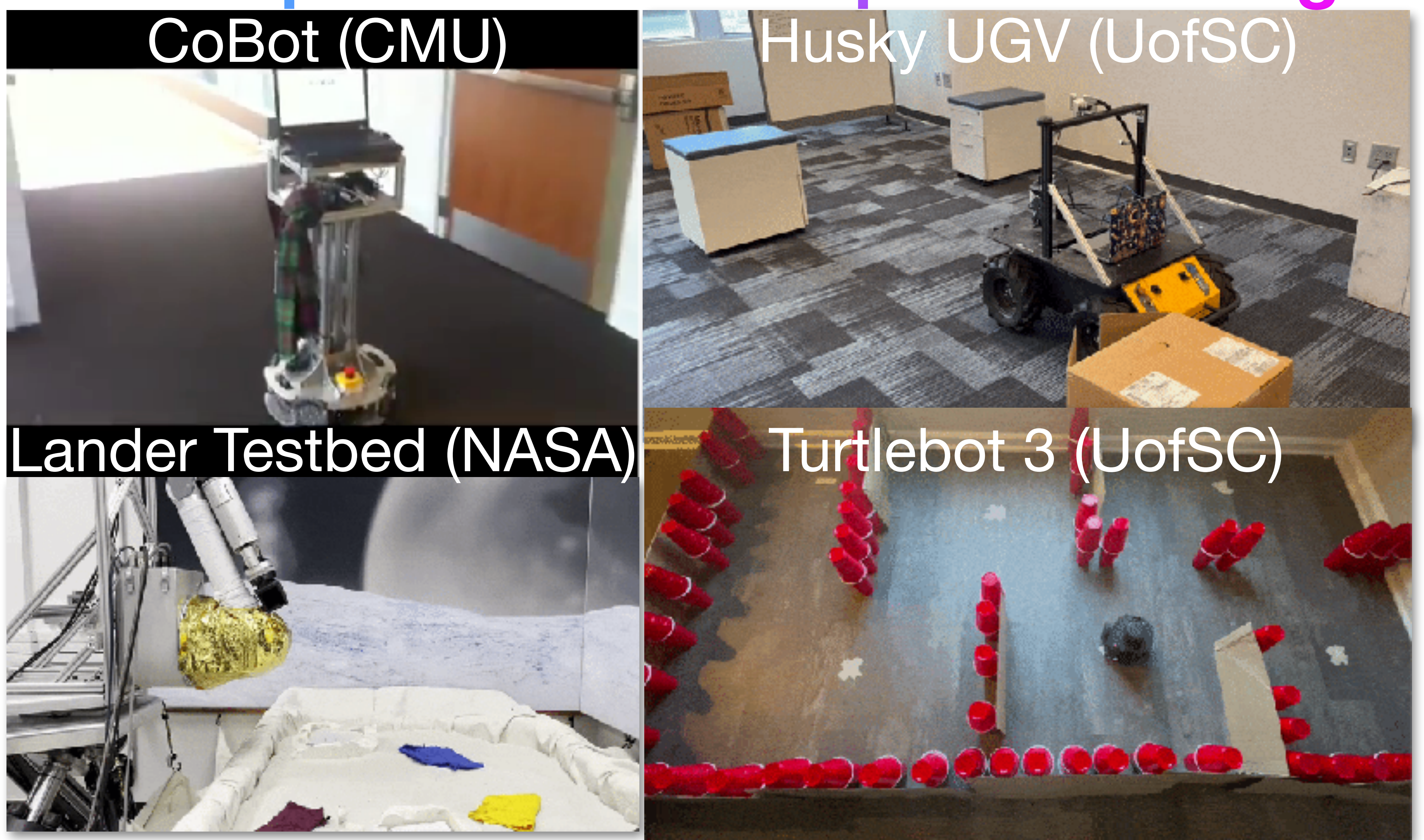
The variability space (design space) of (composed) systems is exponentially increasing



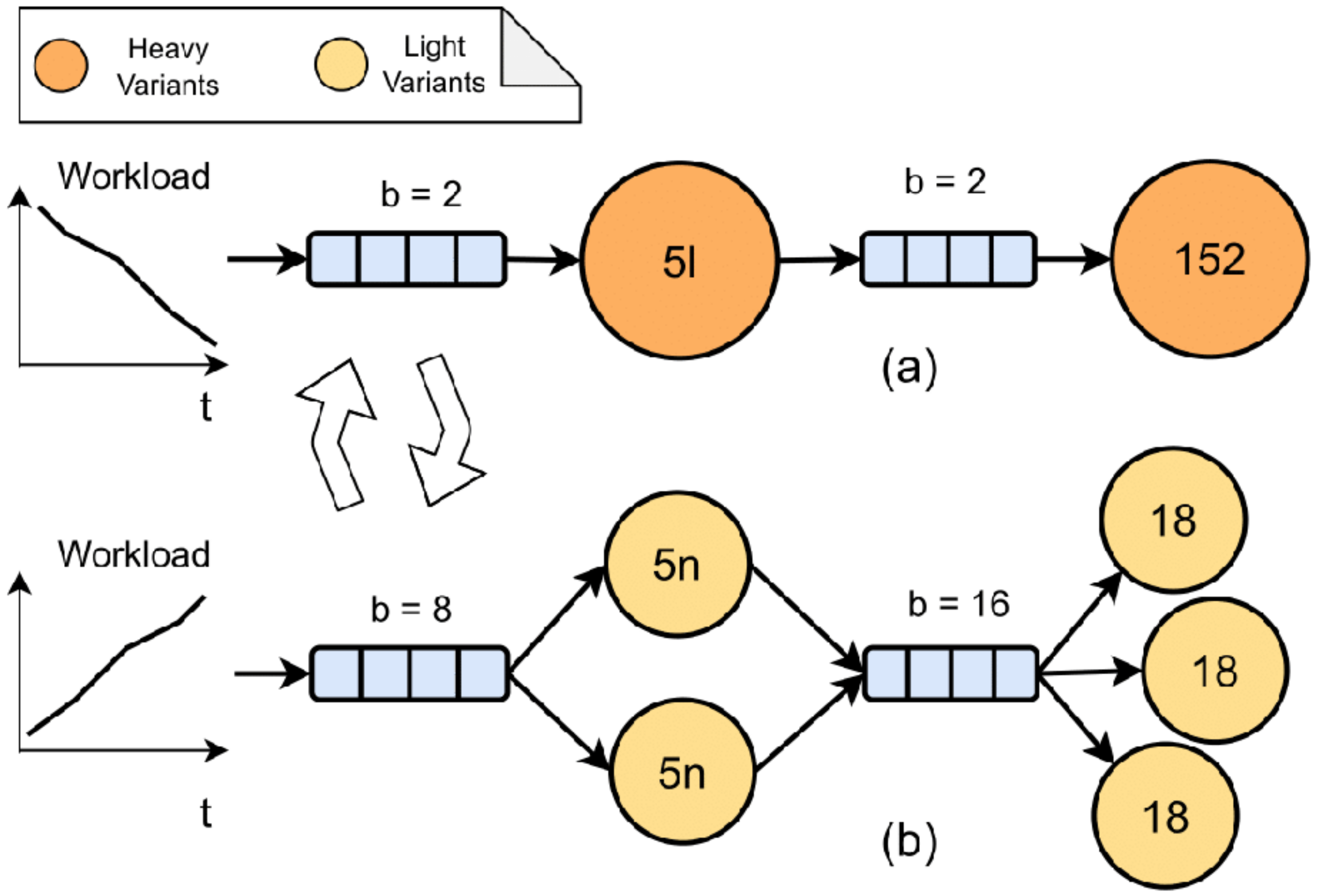
Performance goals are competing and users have preferences over these goals



Systems operate in uncertain environments with imperfect and incomplete knowledge



Goal: Enabling users to find the right quality tradeoff

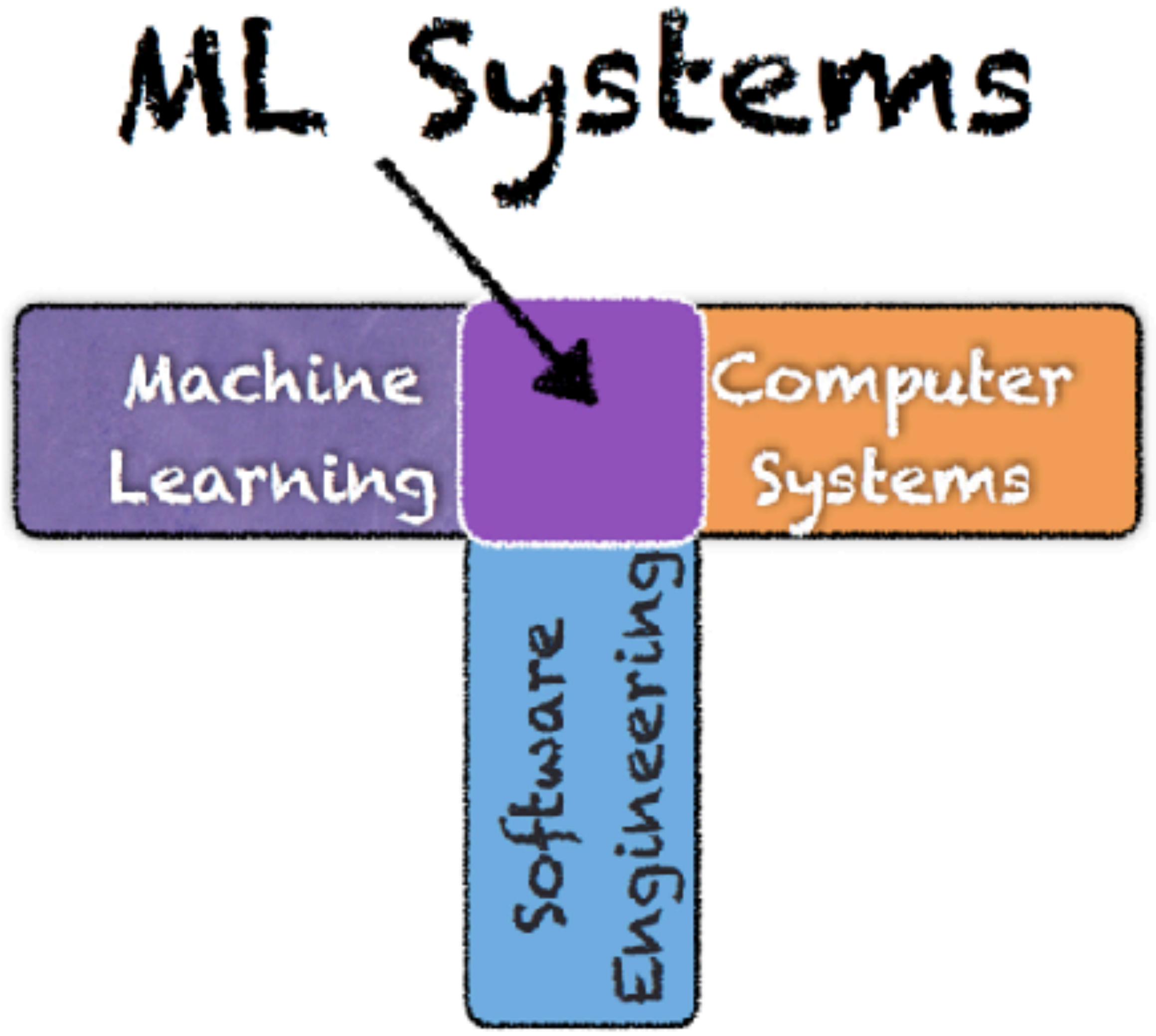


Extension Ideas

Two teams in the CSCE 585 ML Systems course have explored interesting ideas to extend and build on top of IPA infrastructure!

CSCE 585 LECTURES PROJECTS RESOURCES POLICIES PIAZZA

Machine Learning Systems



New to machine learning? Not sure how ML works in production? Interested to get involved in advanced ML+Systems research? This class is designed for you!

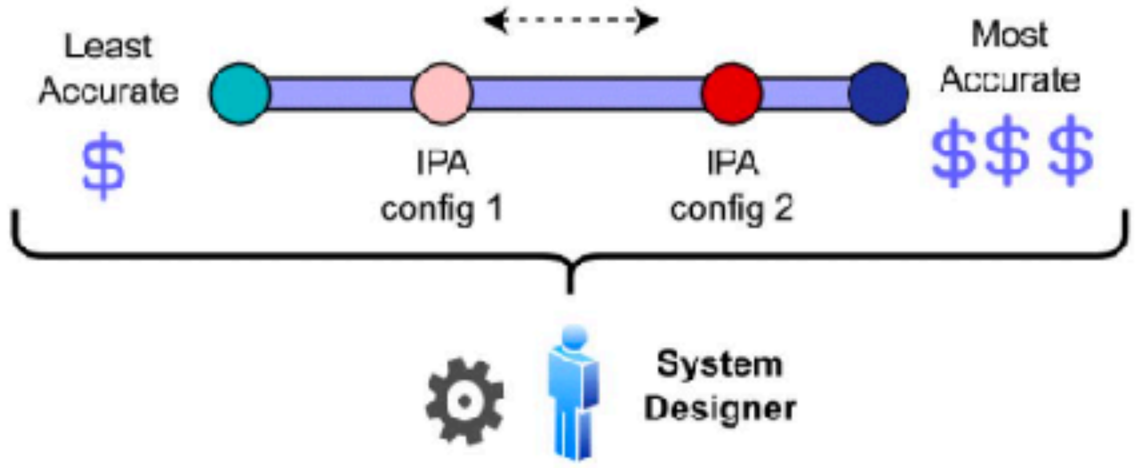
When we talk about Artificial Intelligence (AI) or Machine Learning (ML), we typically refer to a technique, a model, or an algorithm that gives the computer systems the ability to learn and to reason with data. However, there is a lot more to ML than just implementing an algorithm or a technique. In this course, we will learn the fundamental differences between AI/ML as a model versus AI/ML as a system in production.

Course Website: https://pooyanjamshidi.github.io/mls/

UNIVERSITY OF South Carolina

Considering Energy Consumption in IPA Towards Sustainable AI

Regan Willis, Chase Bryson, Osasuyi Agho



https://github.com/csce585-mlsystems/Sustainable-IPA

IPA-Ext

Sabah S. Anis
Computer Science
ML Engineer

Misagh Soltani
Computer Science
ML Research Scientist,
ML Engineer

Xeerak Muhammad
Computer Science
ML Engineer, Scribe, Team Lead

https://github.com/csce585-mlsystems/ipa-ext https://github.com/pooyanjamshidi/modular-composed-ai-systems/tree/main/talks