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

Mehran Salmani, Saeid Ghafouri, **Alireza Sanaee**, Kamran Razavi

Joseph Doyle, Max Mühlhäuser, Pooyan Jamshidi, Mohsen Sharifi

IUST Iran University of Science and Technology

Queen Mary University of London

TECHNISCHE UNIVERSITÄT DARMSTADT

UNIVERSITY OF South Carolina

“More than 90% of data center compute for ML workload, is used by inference services”

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!

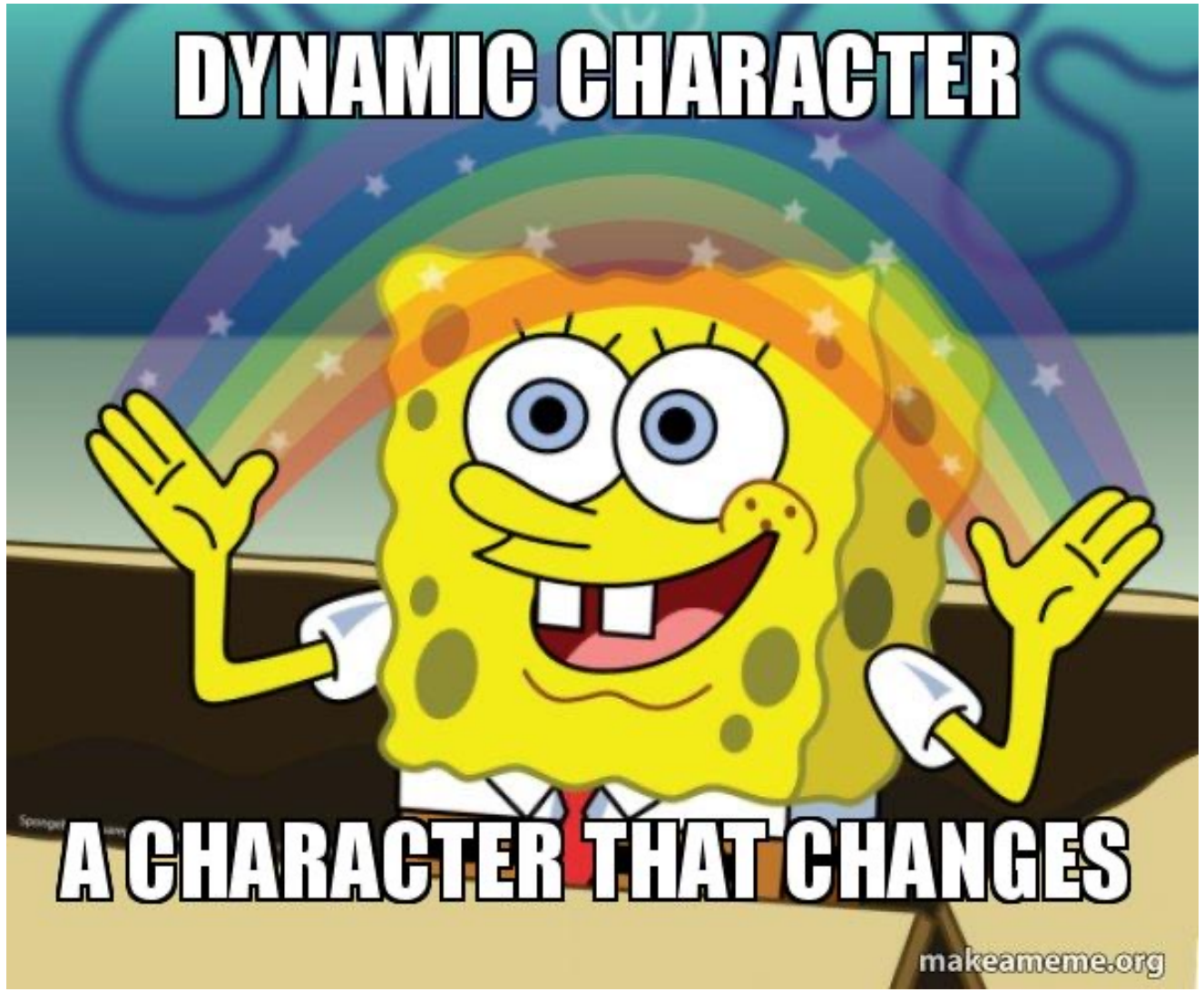
ML inference services have strict & **conflicting** requirements

Highly Responsive!

Cost-Efficient!

Highly Accurate!

More challenge: Dynamic workload

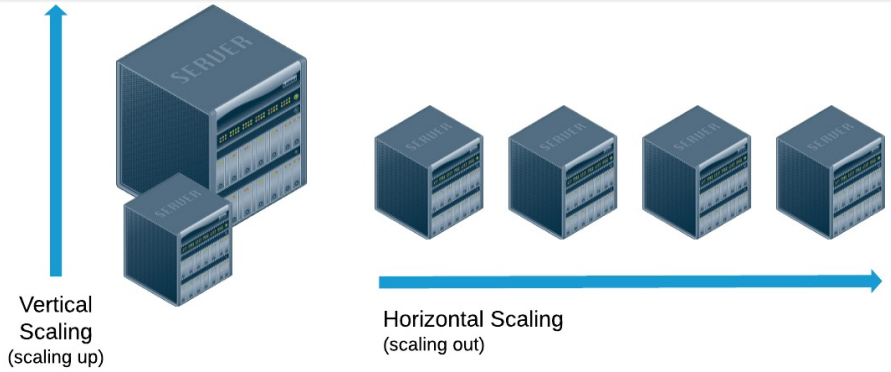


Existing adaptation mechanisms

Resource Scaling

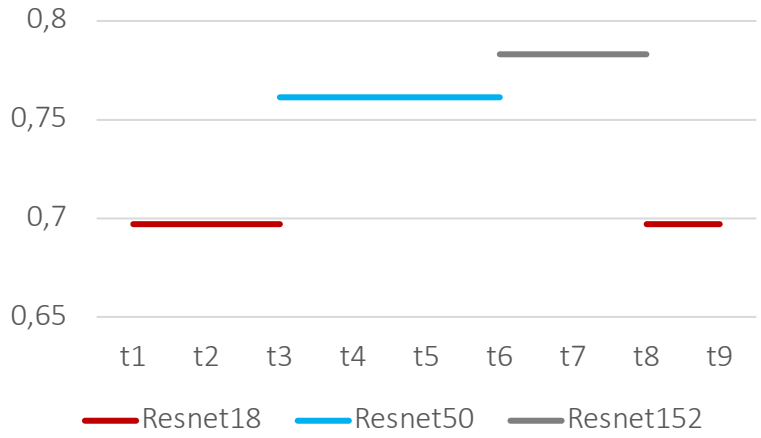
Vertical Scaling (AutoPilot EuroSys'20)

Horizontal Scaling (MArk ATC'19)



Quality Adaptation

Multi Variants (Model-Switching Hotcloud'20)

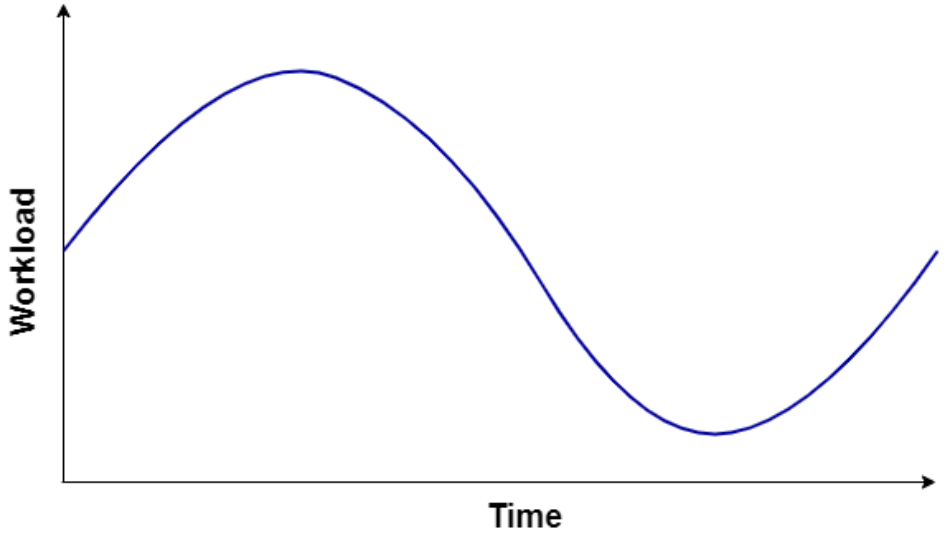


Resource allocation

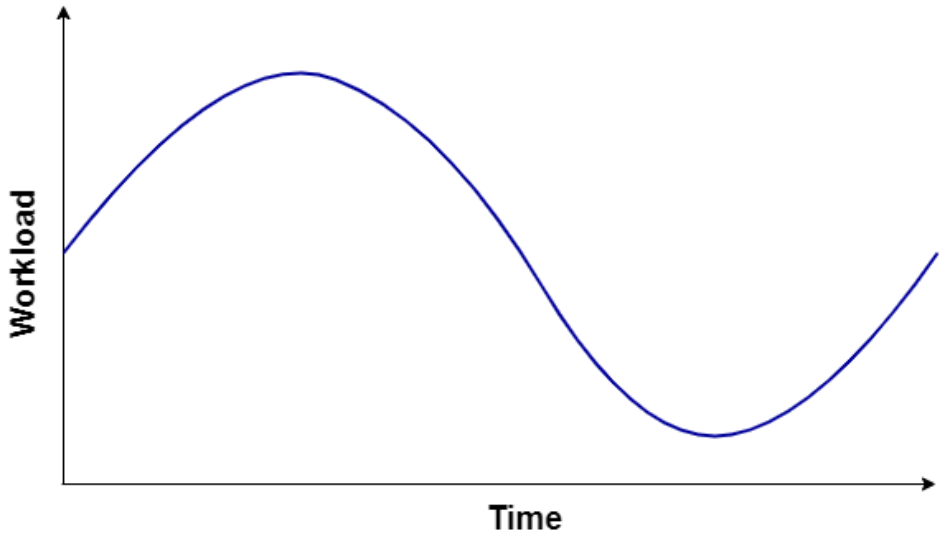
Over Provisioning

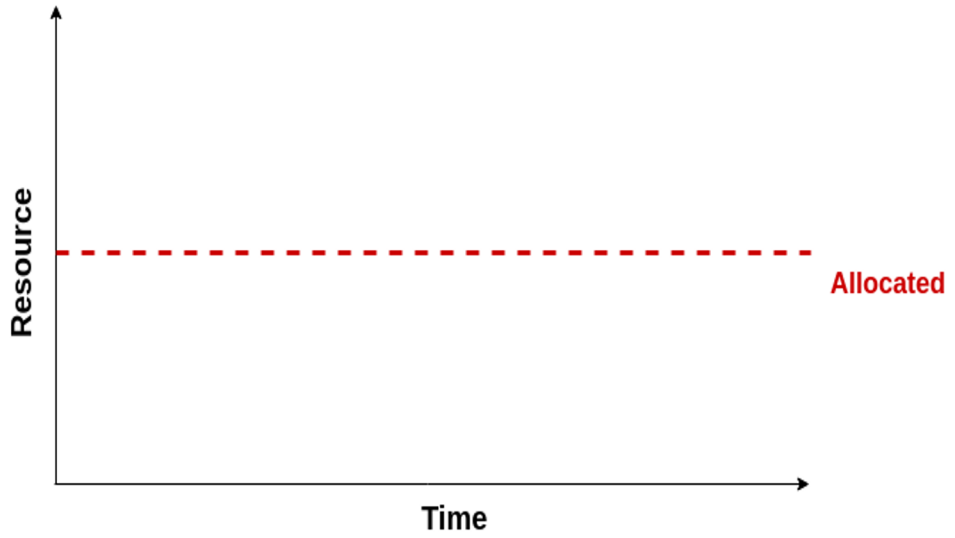
Under Provisioning

Resource allocation

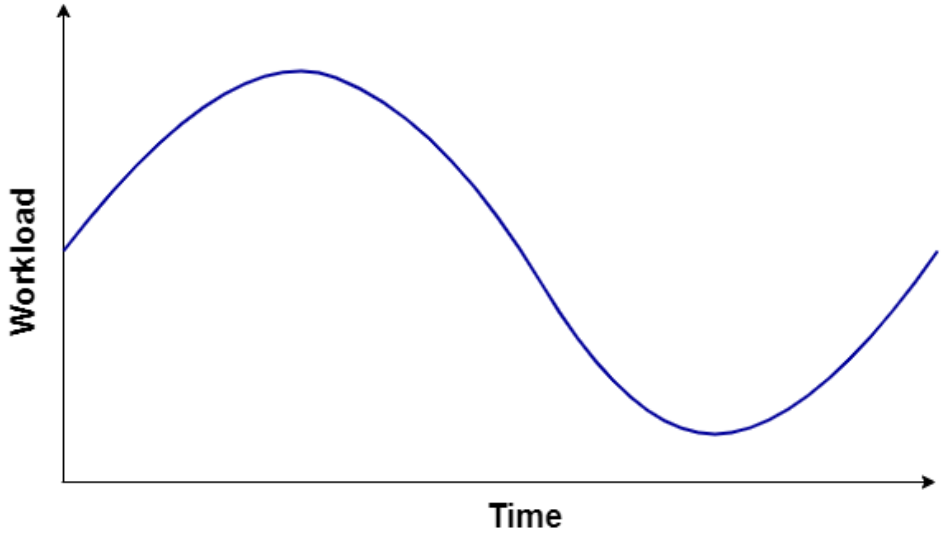


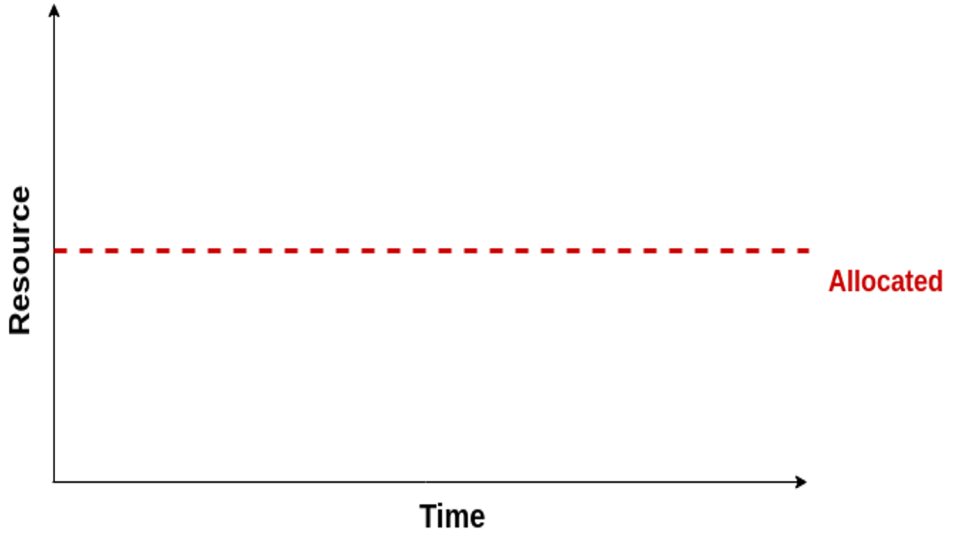
Resource allocation

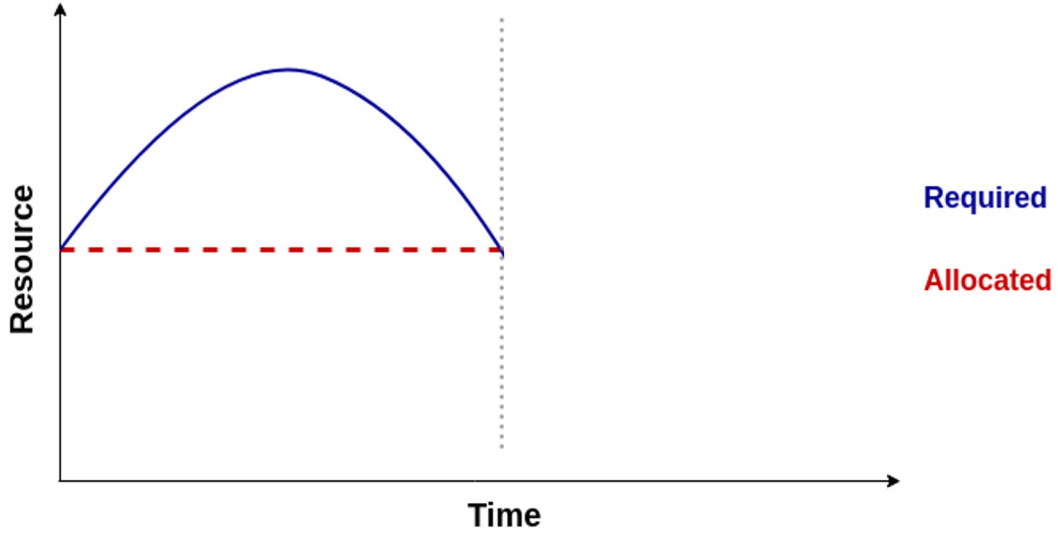




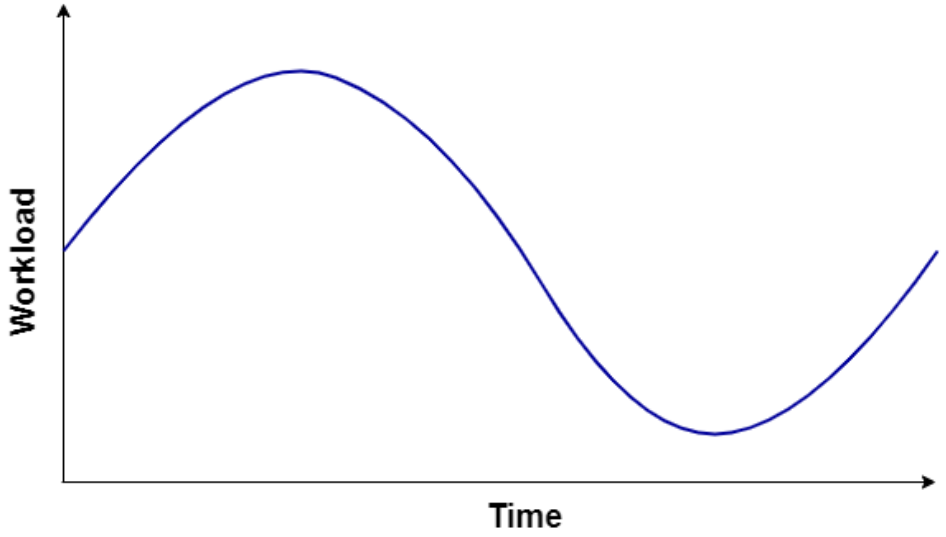
Resource allocation

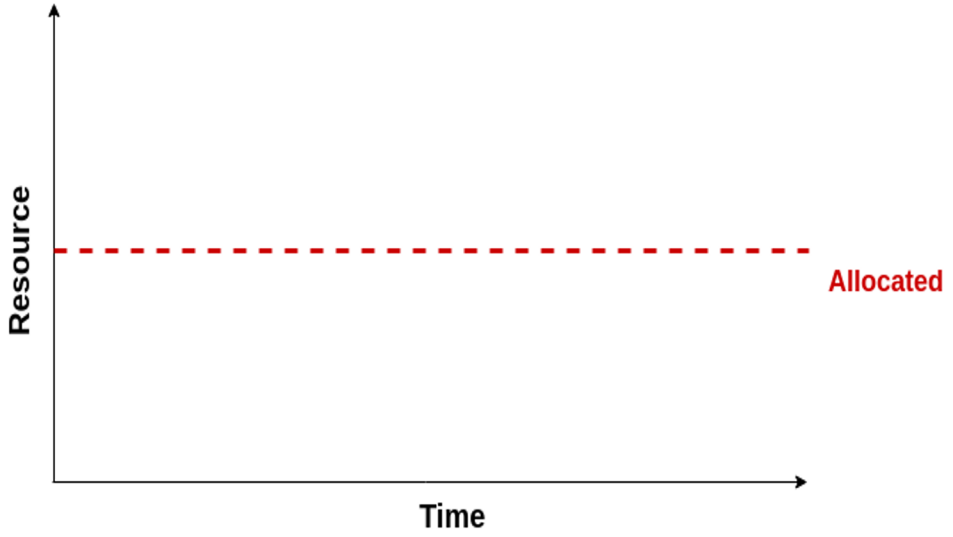


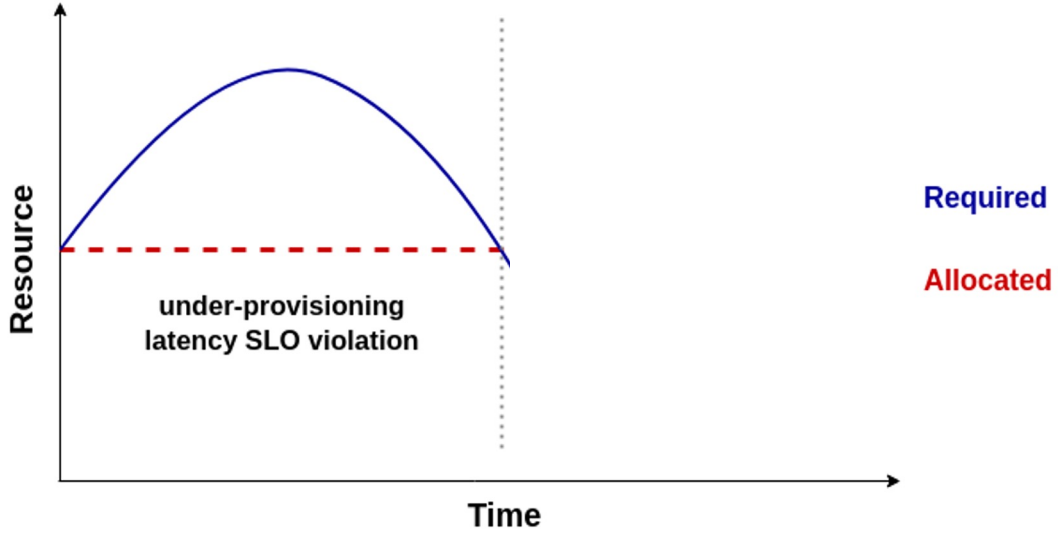




Resource allocation



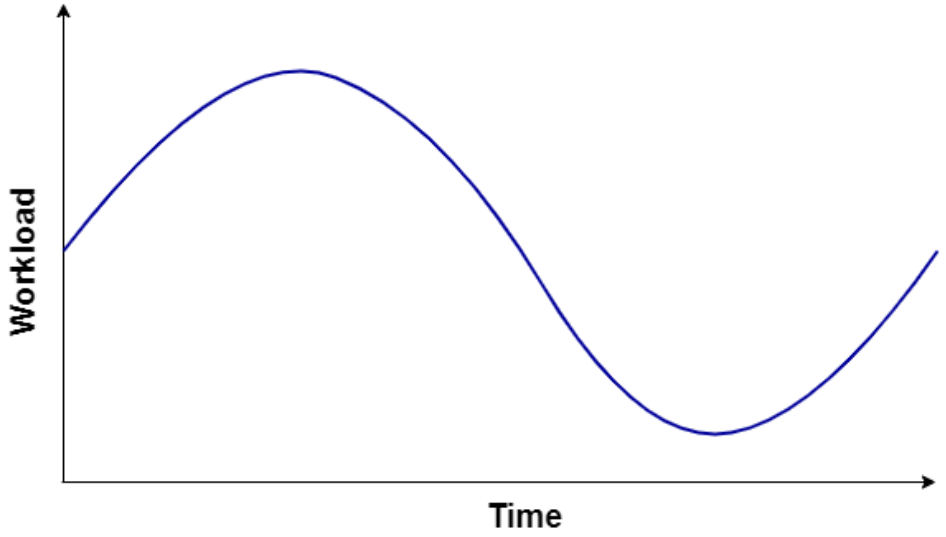


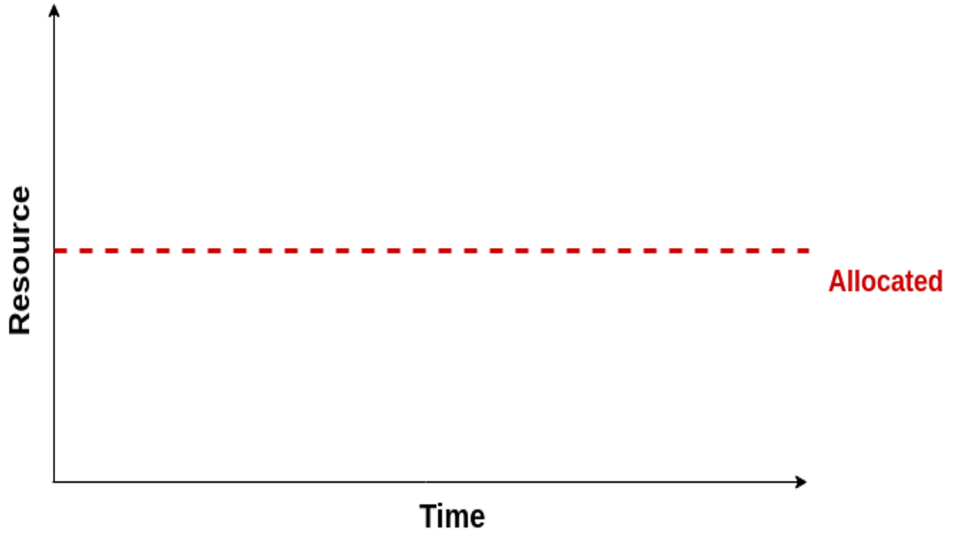


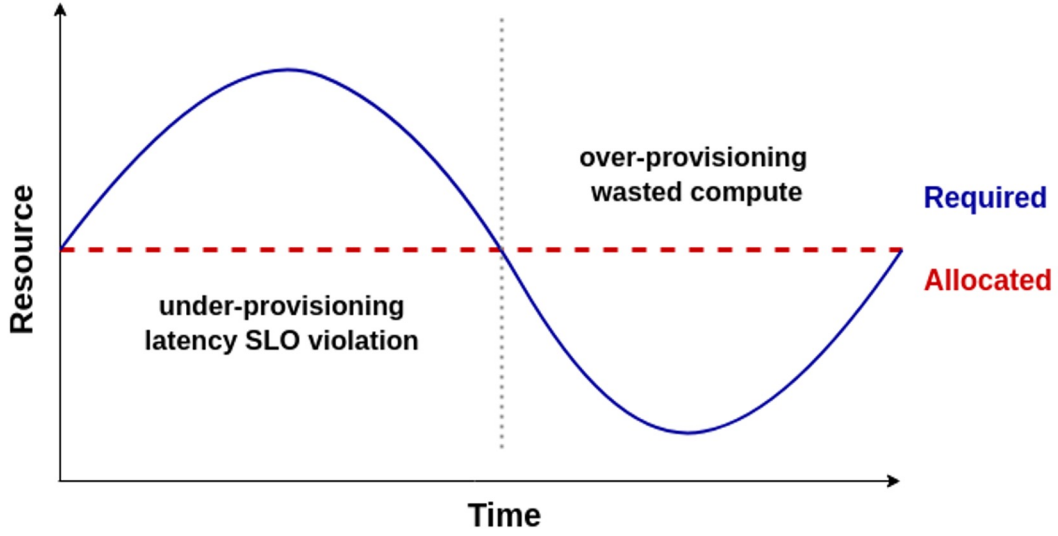
Required

Allocated

Resource allocation





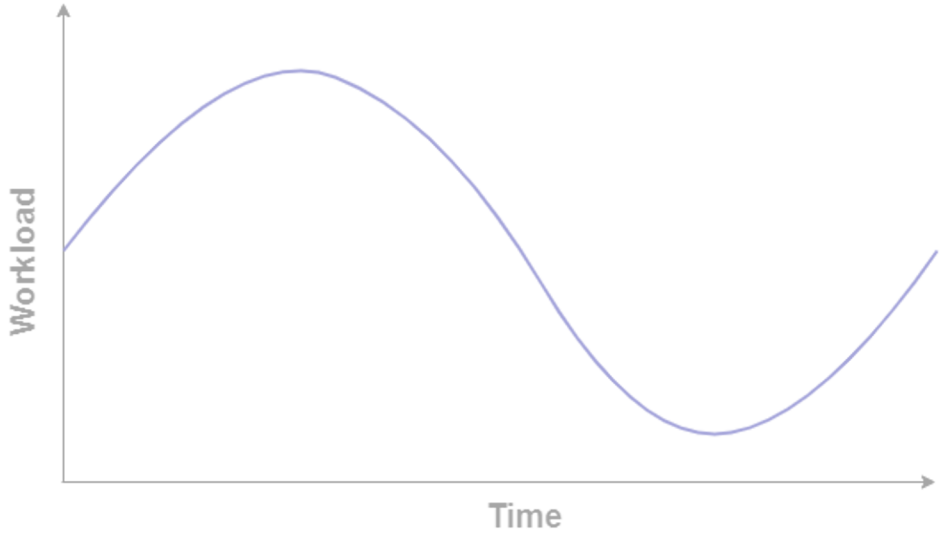


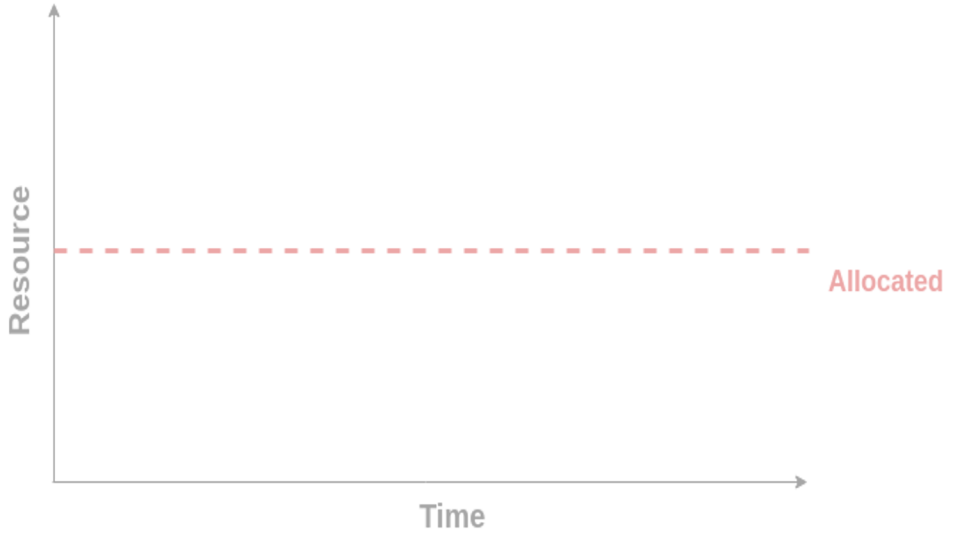
Quality adaptation

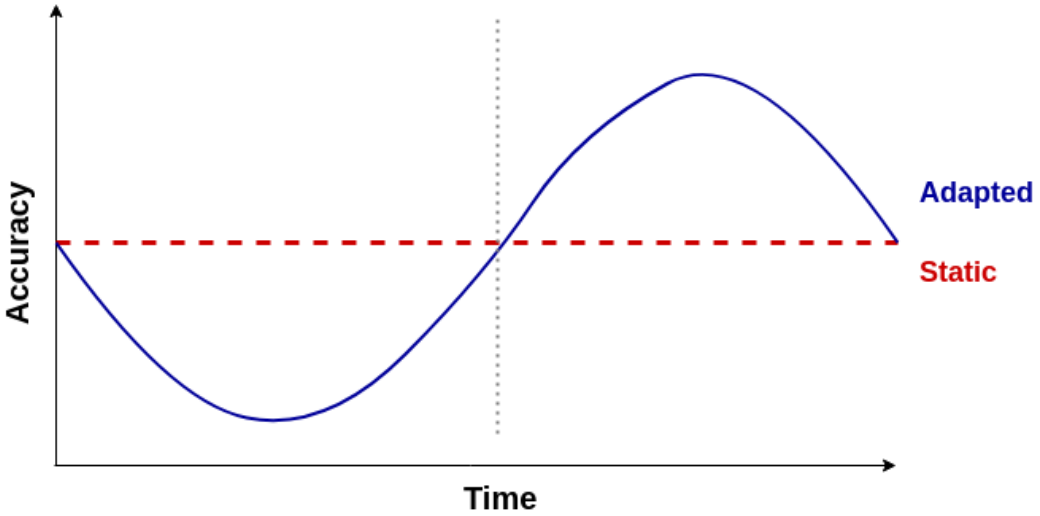
ResNet18: Tiger

ResNet152: Dog

Quality adaptation





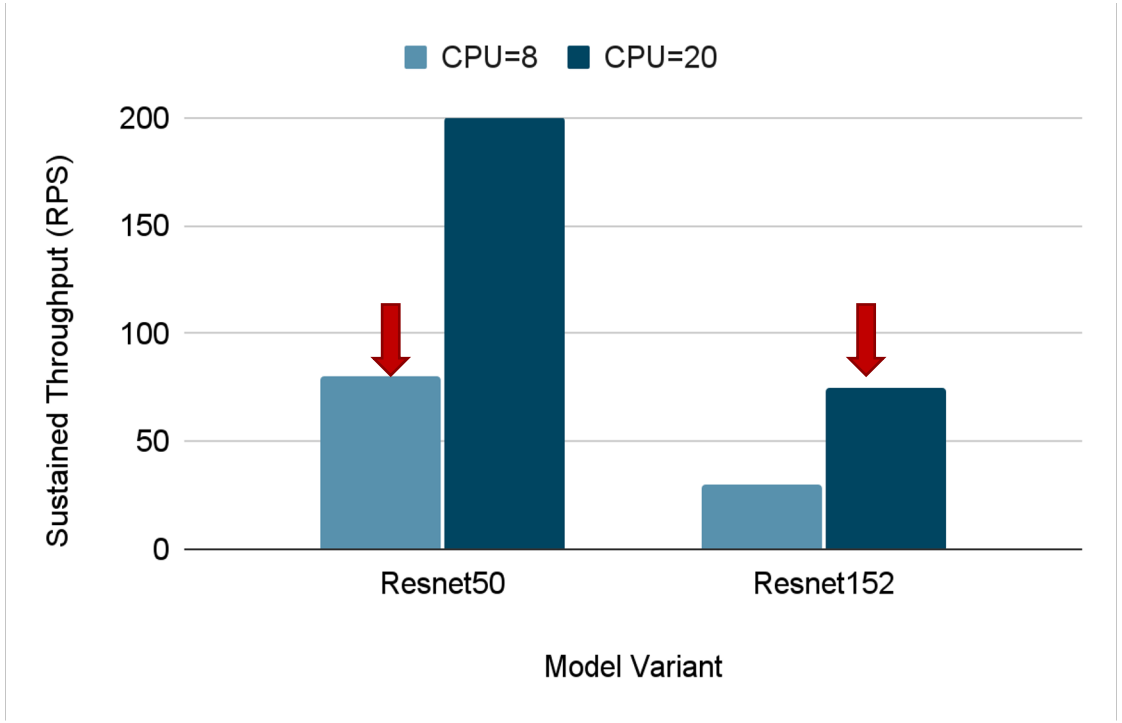


Solution: InfAdapter

InfAdapter is a latency SLO-aware, highly accurate, and cost-efficient inference serving system.

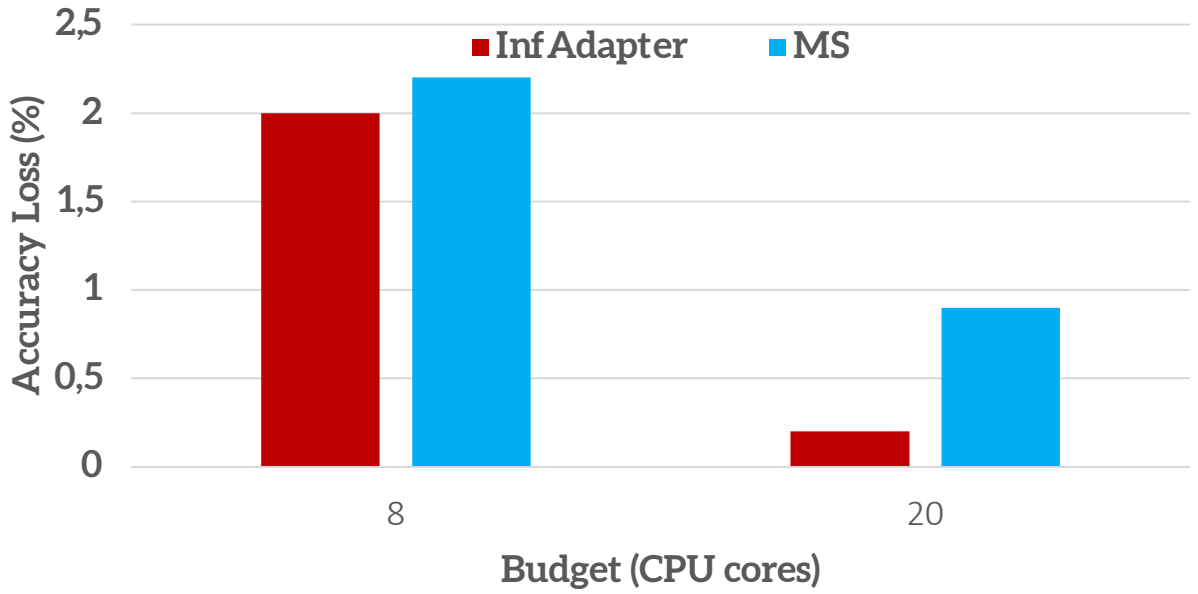
InfAdapter: Why?

Different throughputs with different model variants

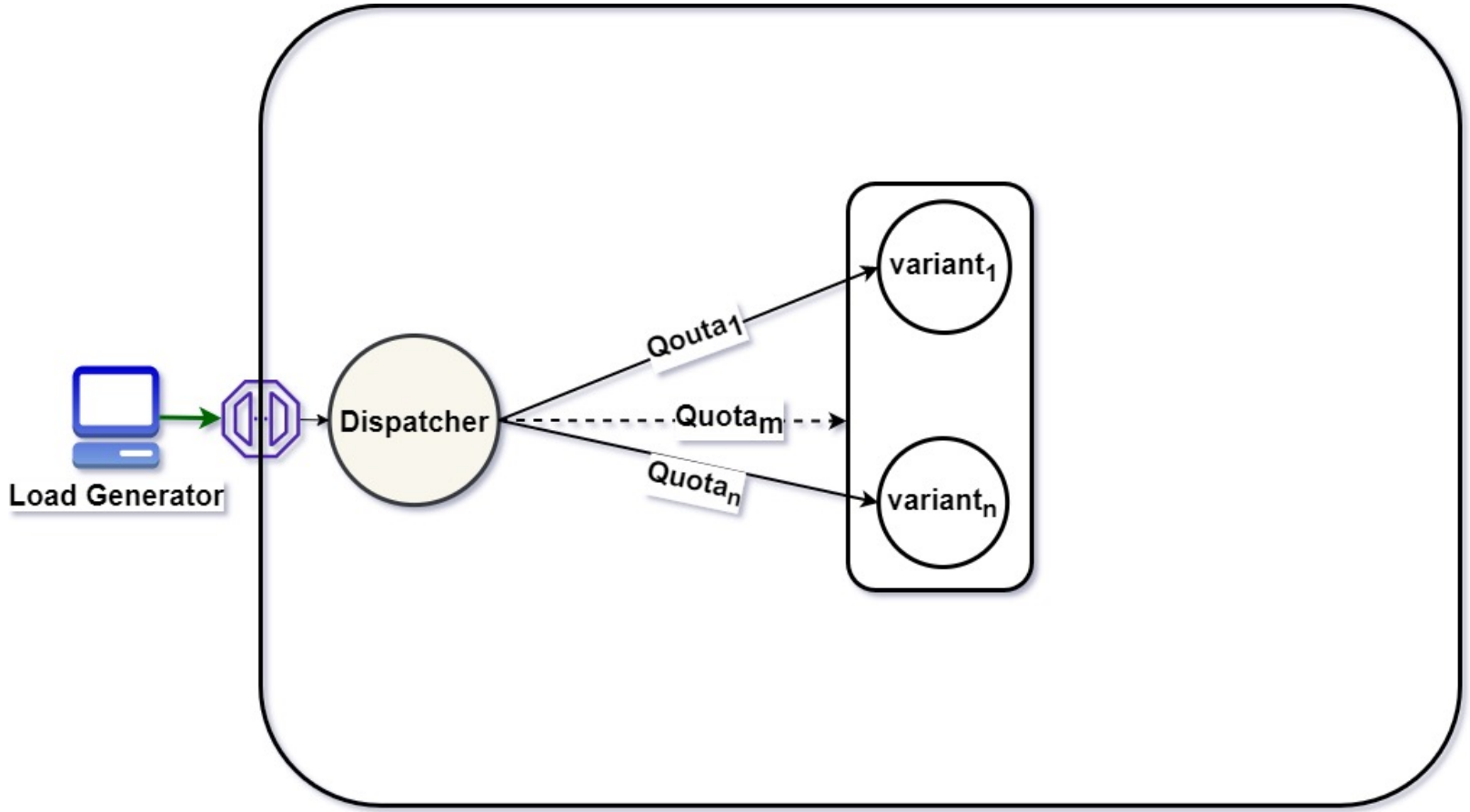


InfAdapter: Why?

Higher average accuracy by using multiple model variants



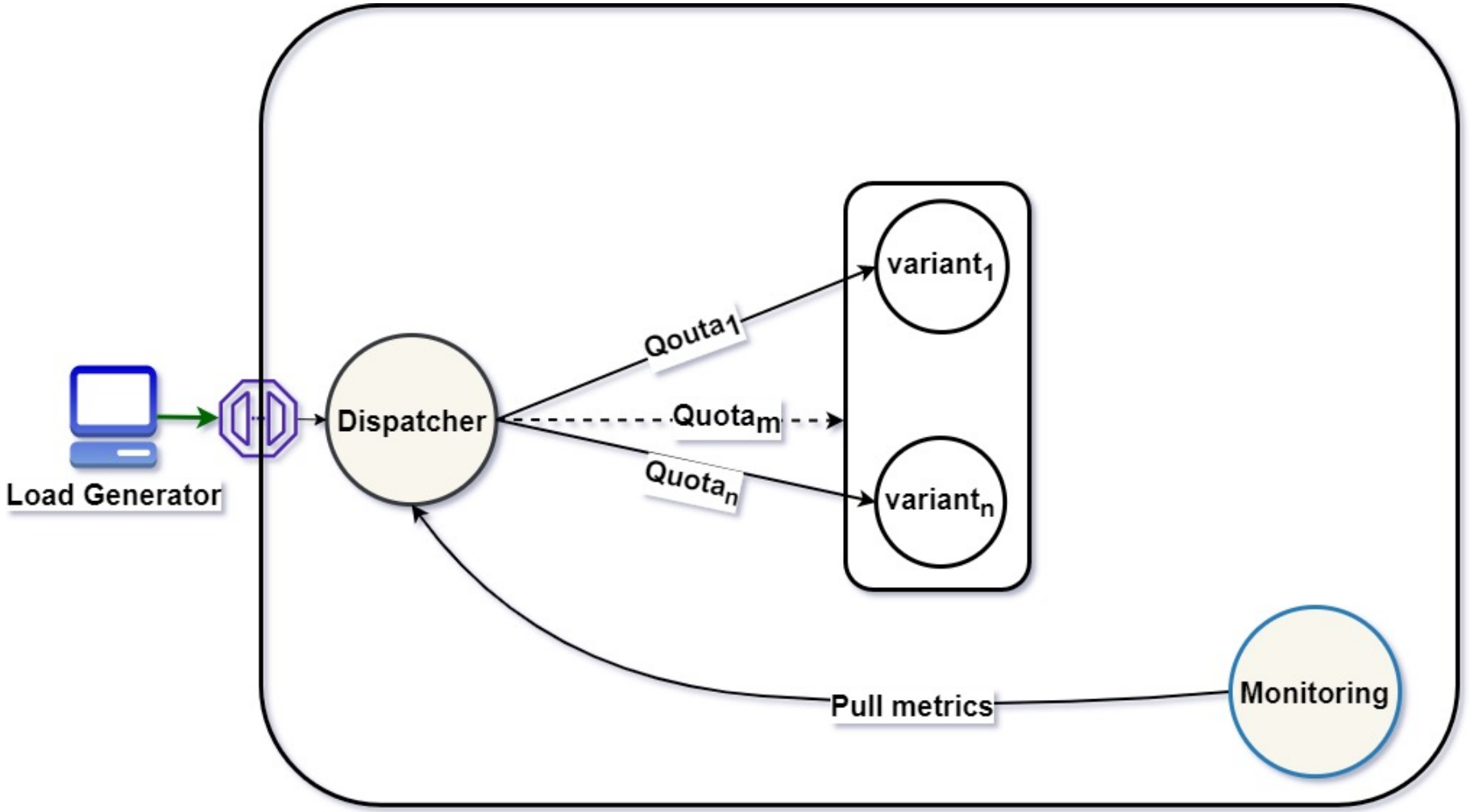
InfAdapter: How?



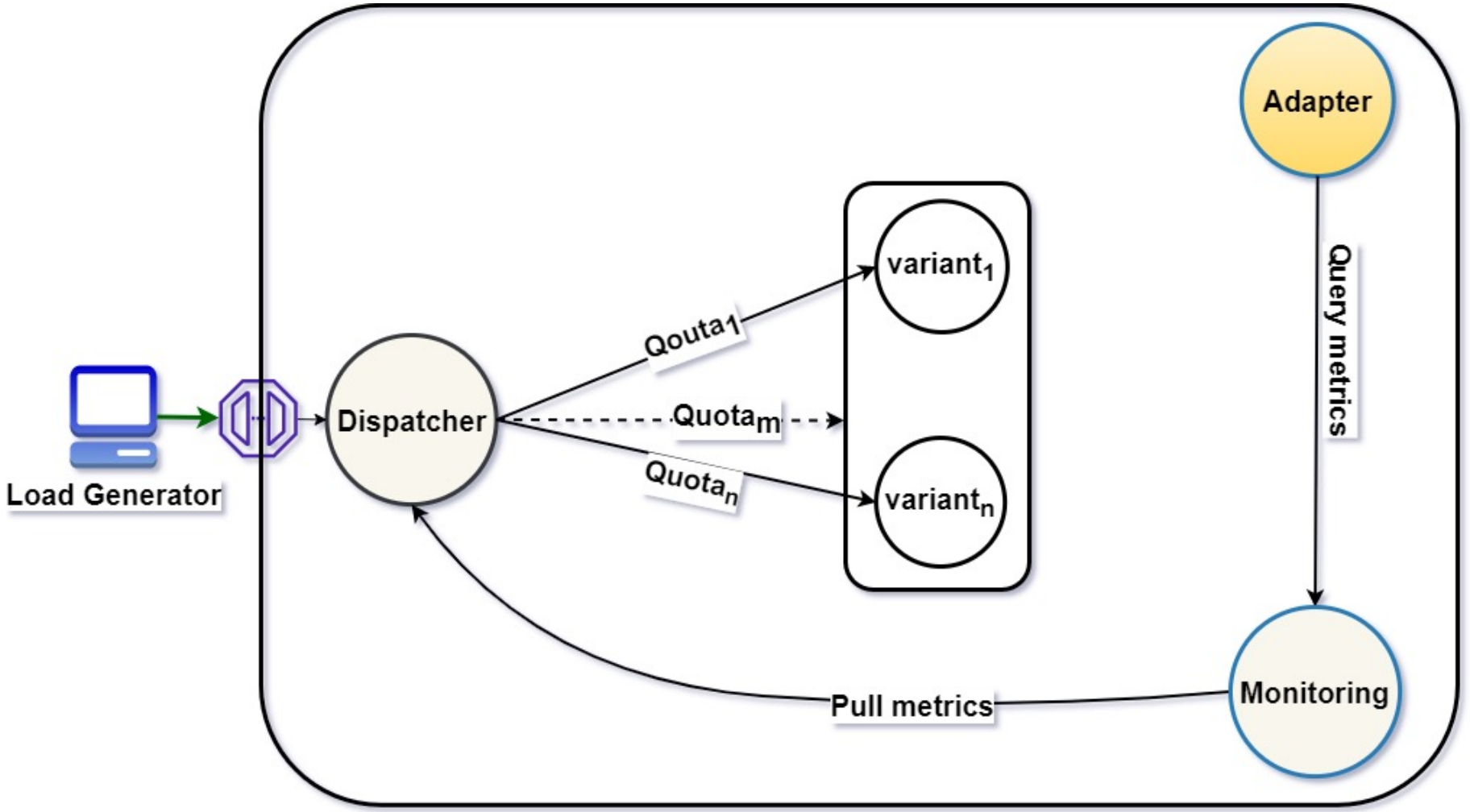
Selecting a subset of model variants, each having its own size

Meeting latency requirement for the predicted workload while maximizing accuracy and minimizing cost

InfAdapter: Design



InfAdapter: Design



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

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

$$\text{subject to} \quad \lambda \leq \sum_{m \in M} th_m(n_m),$$

Supporting incoming workload

$$\lambda_m \leq th_m(n_m)$$

$$p_m(n_m) \leq L, \forall m \in M,$$

$$RC \leq B,$$

$$n_m \in \mathbb{W}, \forall m \in M.$$

InfAdapter: Formulation

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

$$\text{subject to} \quad \lambda \leq \sum_{m \in M} th_m(n_m),$$

Supporting incoming workload

$$\lambda_m \leq th_m(n_m)$$

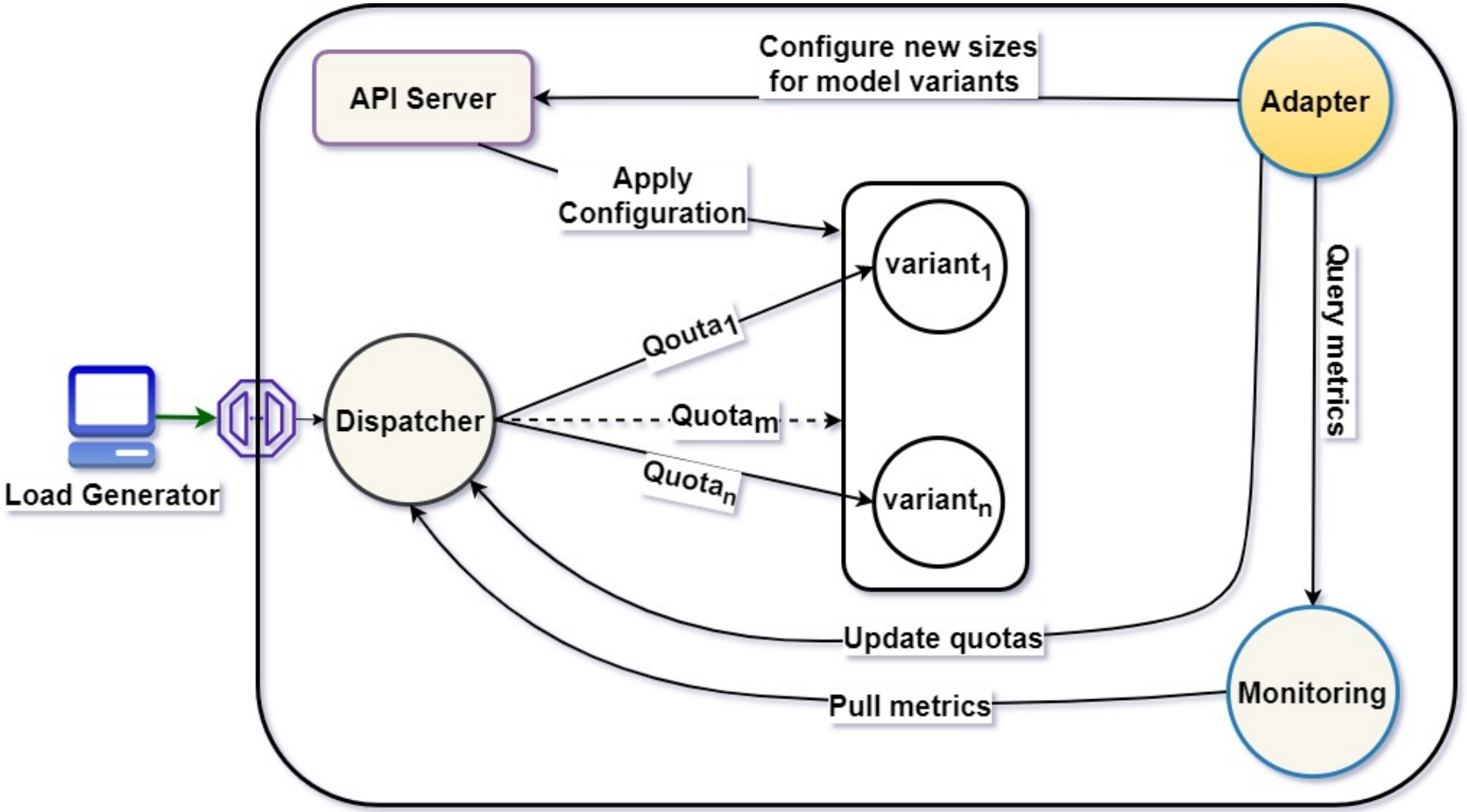
Guaranteeing end-to-end latency

$$p_m(n_m) \leq L, \forall m \in M,$$

$$RC \leq B,$$

$$n_m \in \mathbb{W}, \forall m \in M.$$

InfAdapter: Design



InfAdapter: Experimental evaluation setup

Twitter-trace sample (2022-08)

Baselines

- Kubernetes VPA and adapted Model-Switching

Used models

- Resnet18, Resnet34, Resnet50, Resnet101, Resnet152

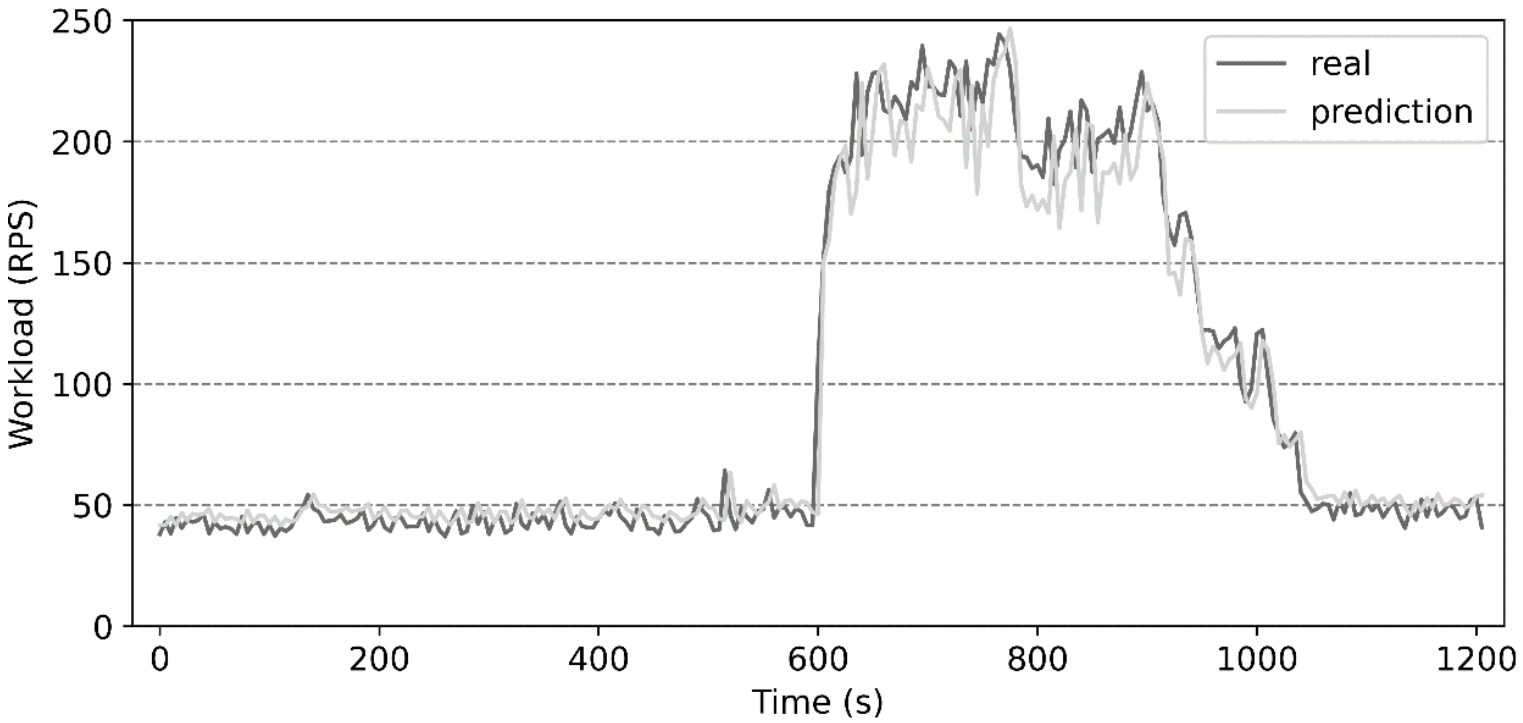
Interval adaptation

- 30 seconds

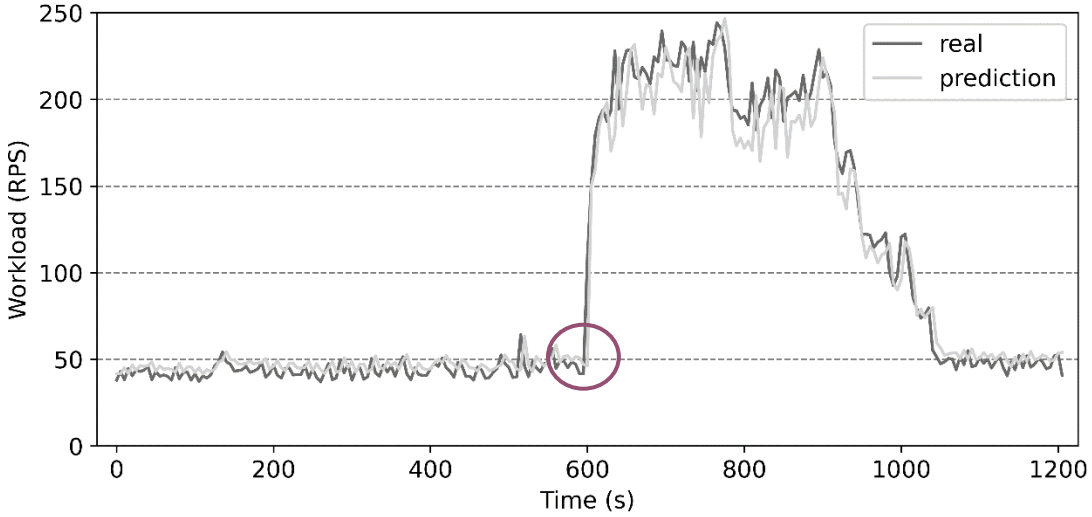
A Kubernetes cluster of 2 computing nodes

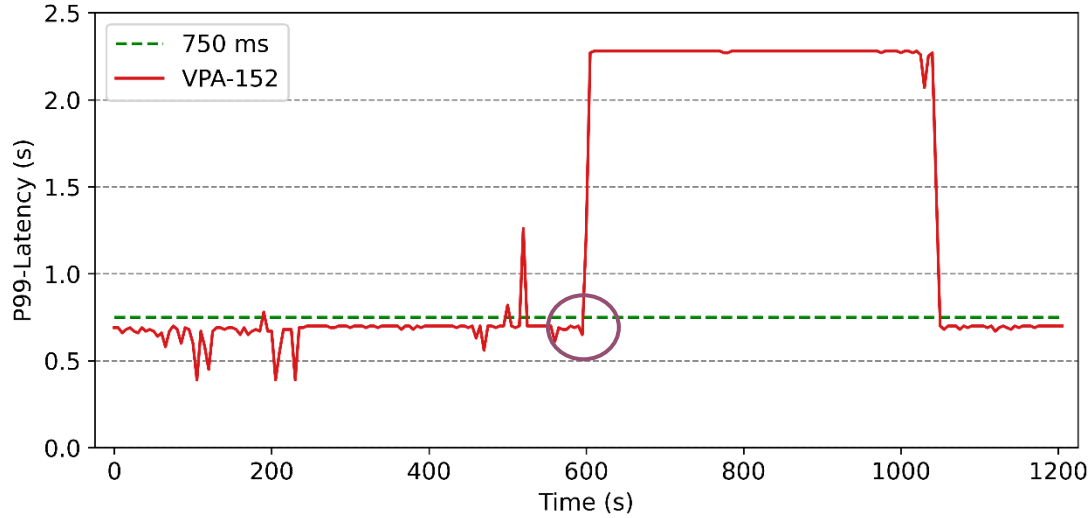
- 48 Cores, 192 GiB RAM

Workload Pattern

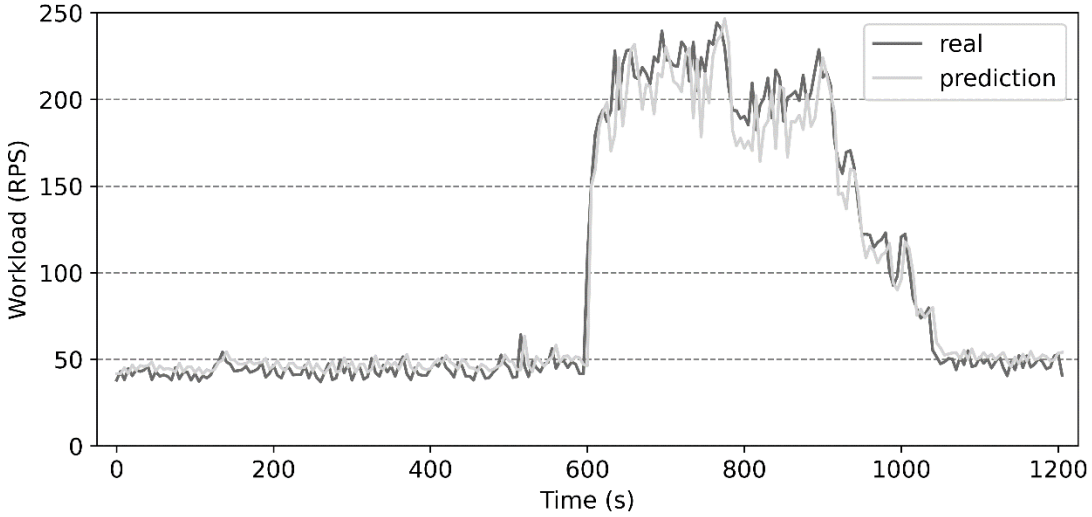


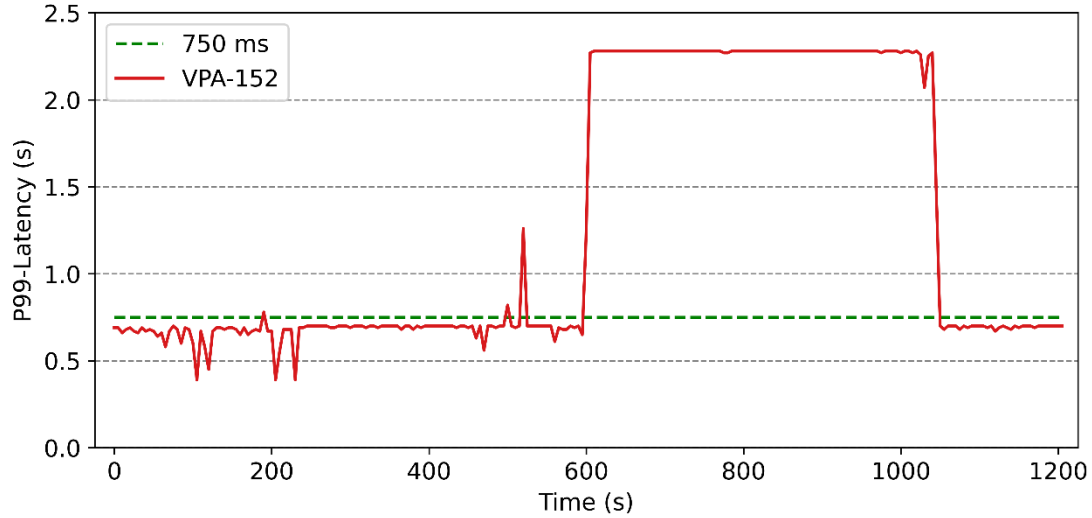
InfAdapter: P99-Latency evaluation



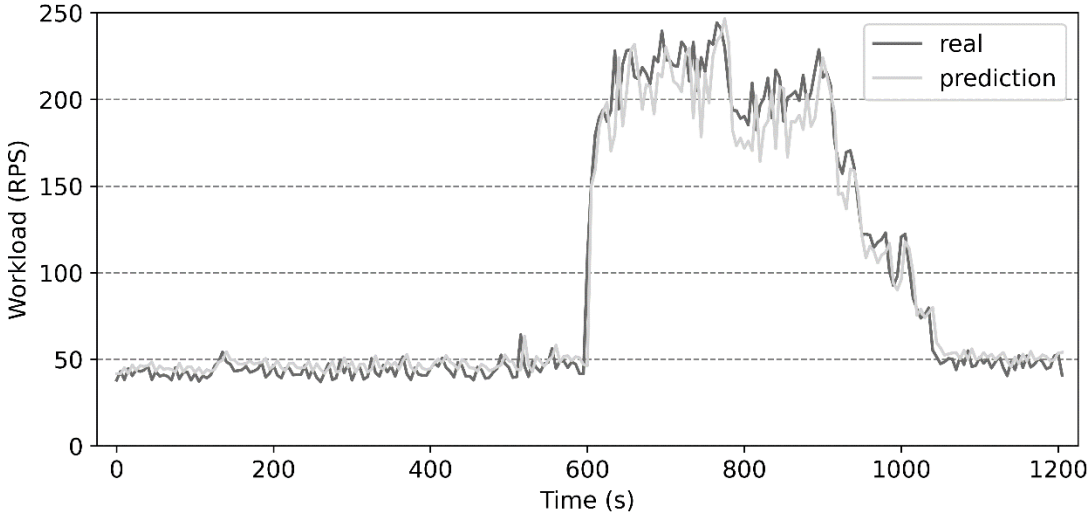


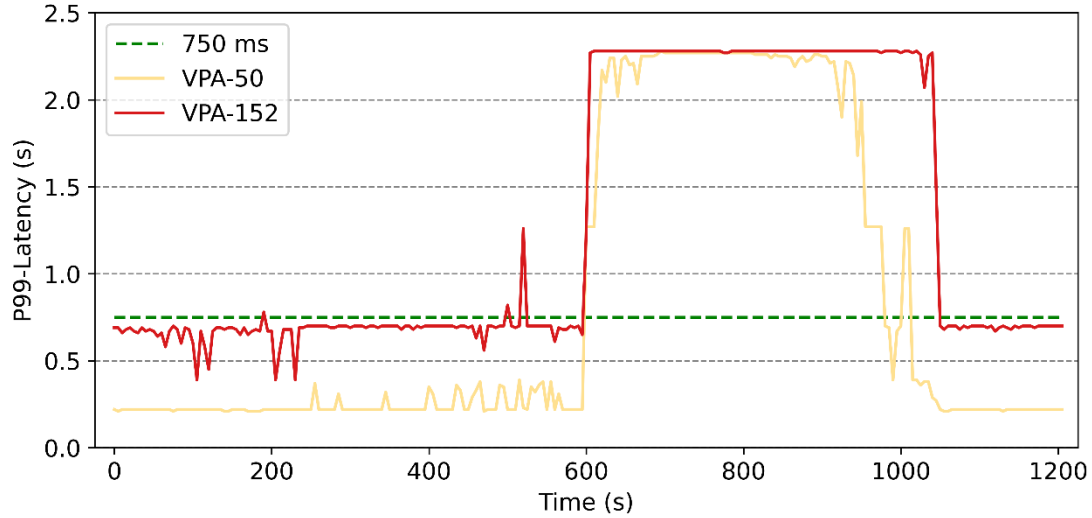
InfAdapter: P99-Latency evaluation



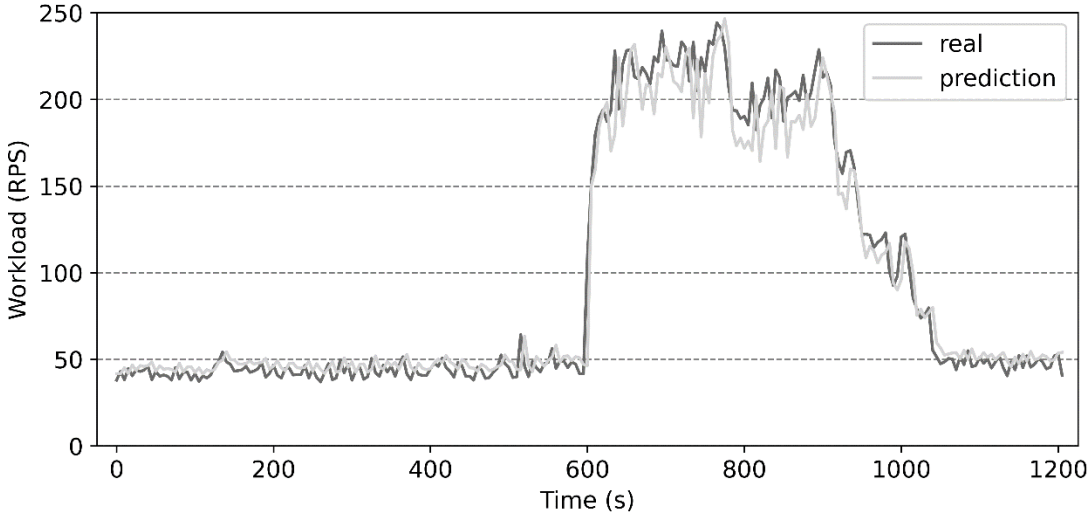


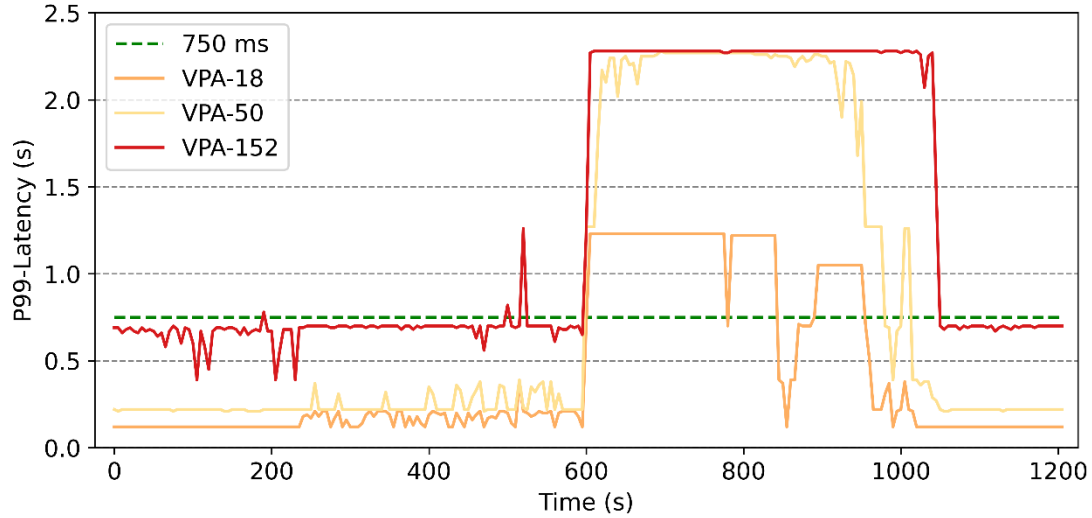
InfAdapter: P99-Latency evaluation



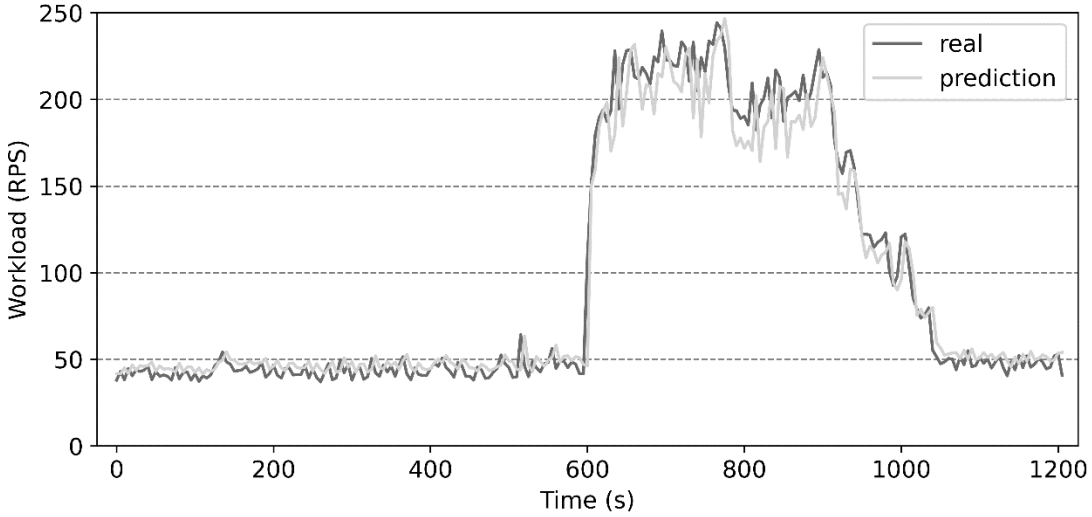


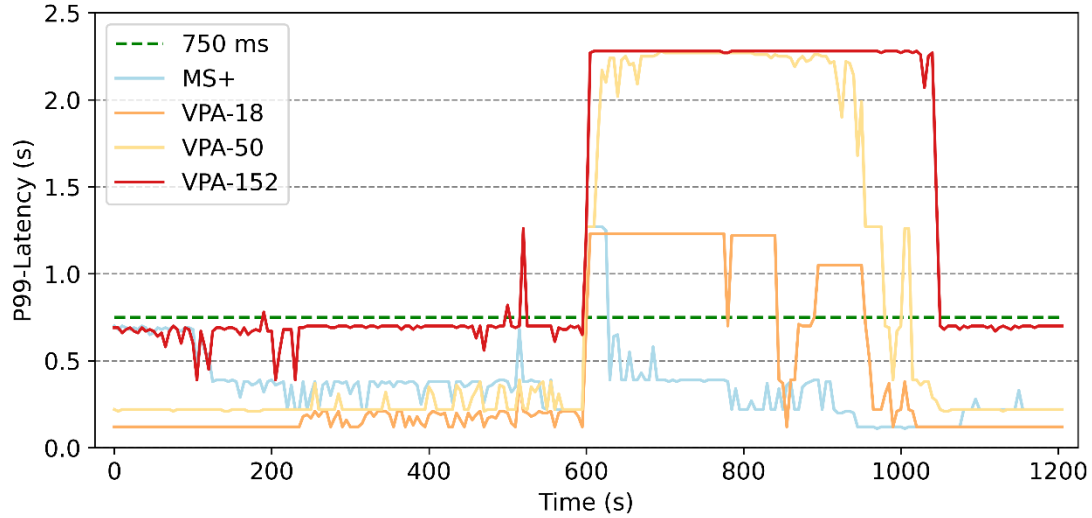
InfAdapter: P99-Latency evaluation



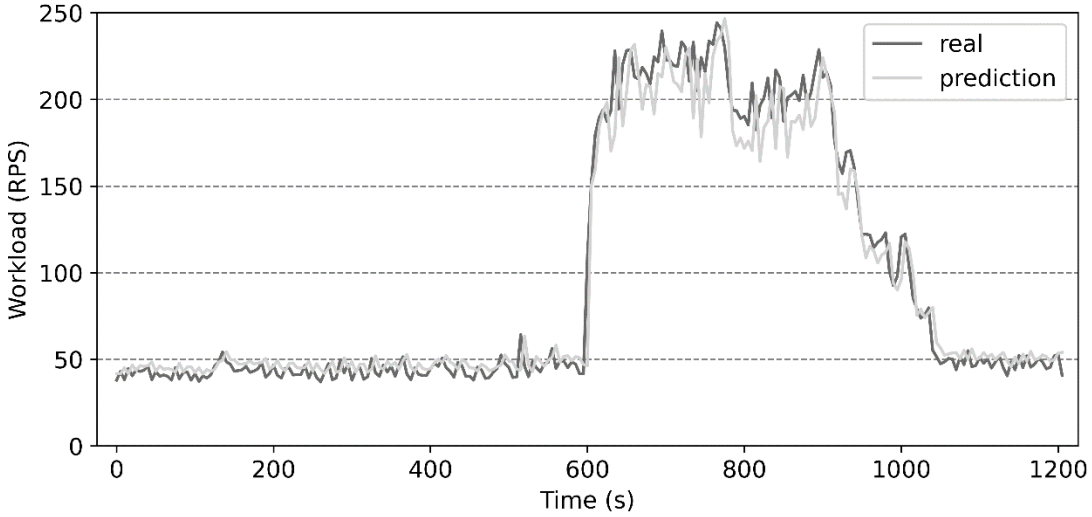


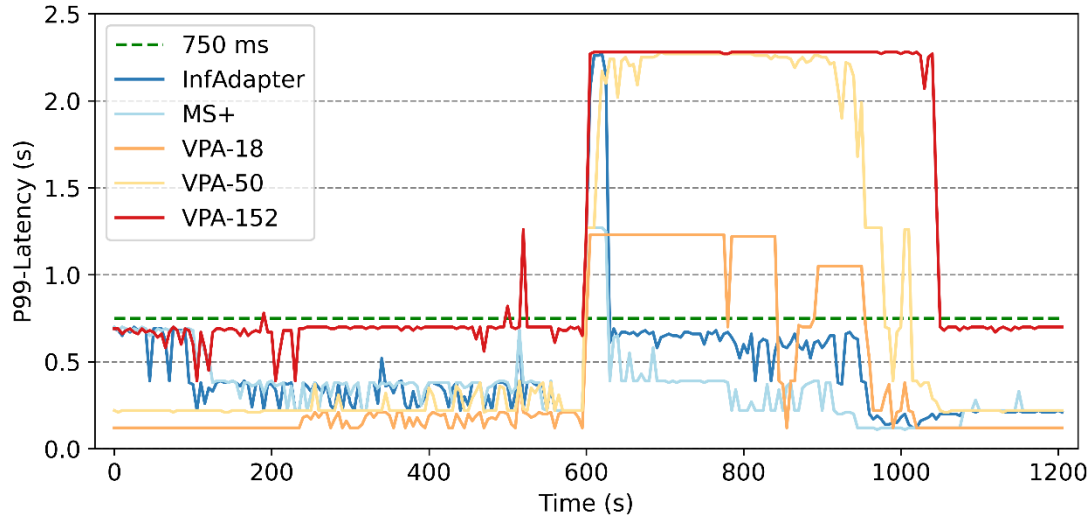
InfAdapter: P99-Latency evaluation



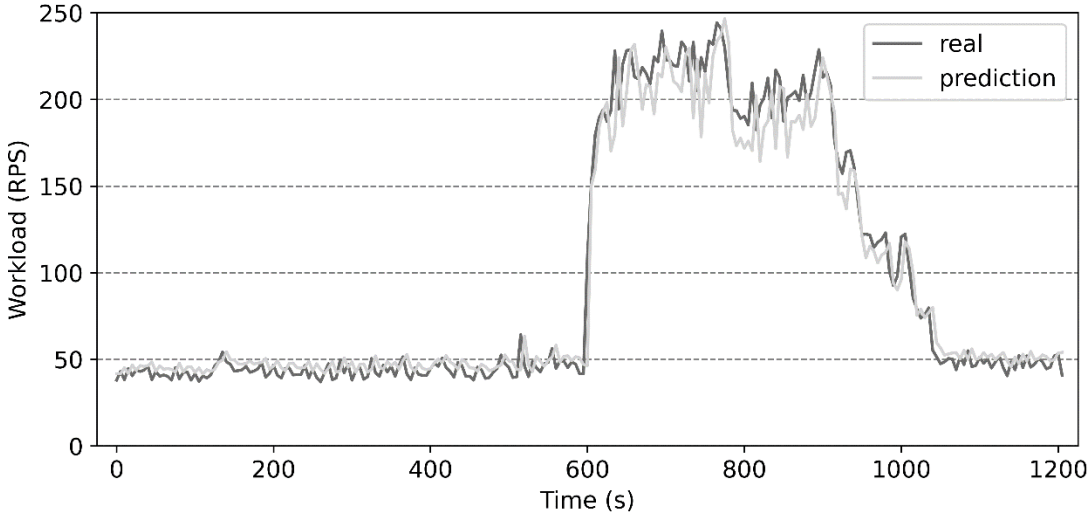


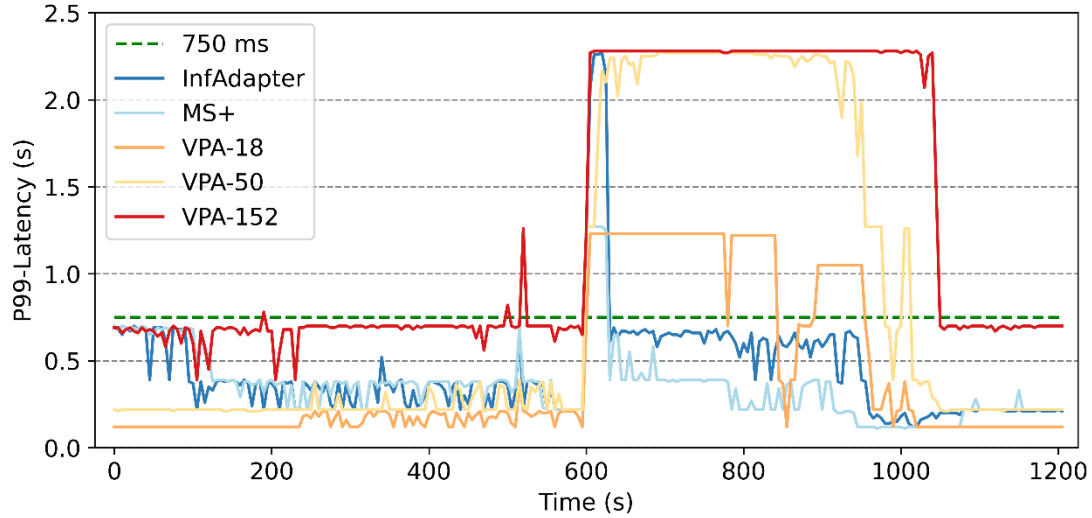
InfAdapter: P99-Latency evaluation



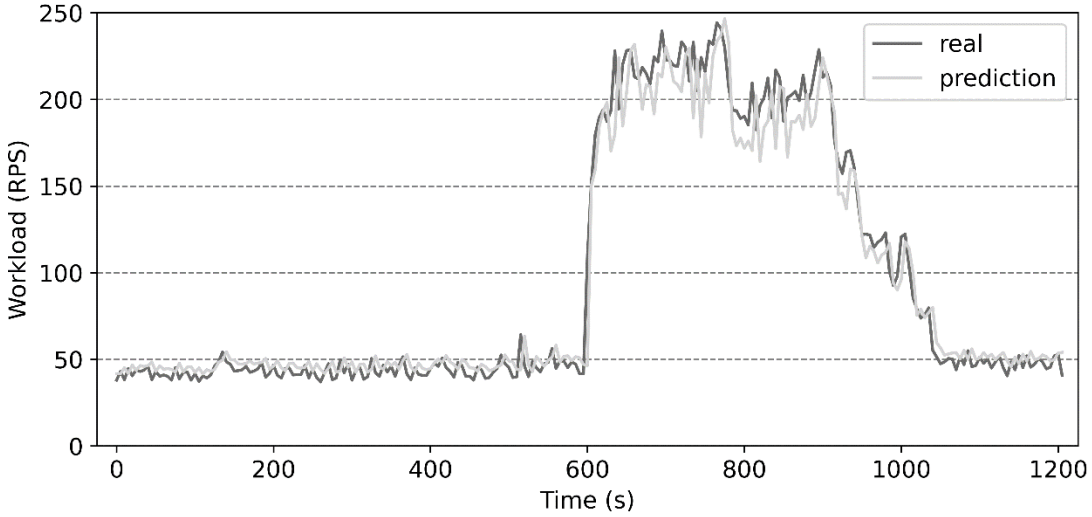


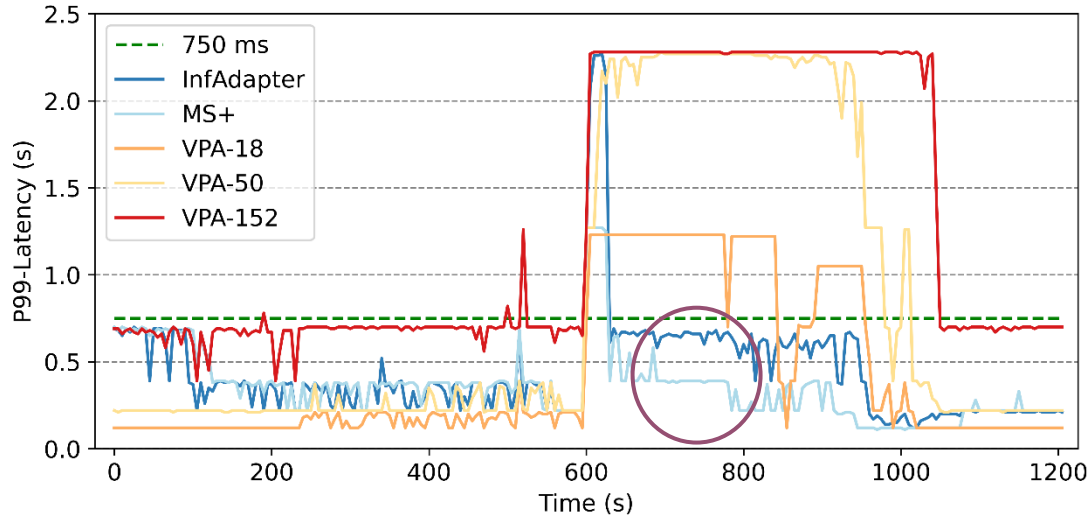
InfAdapter: P99-Latency evaluation



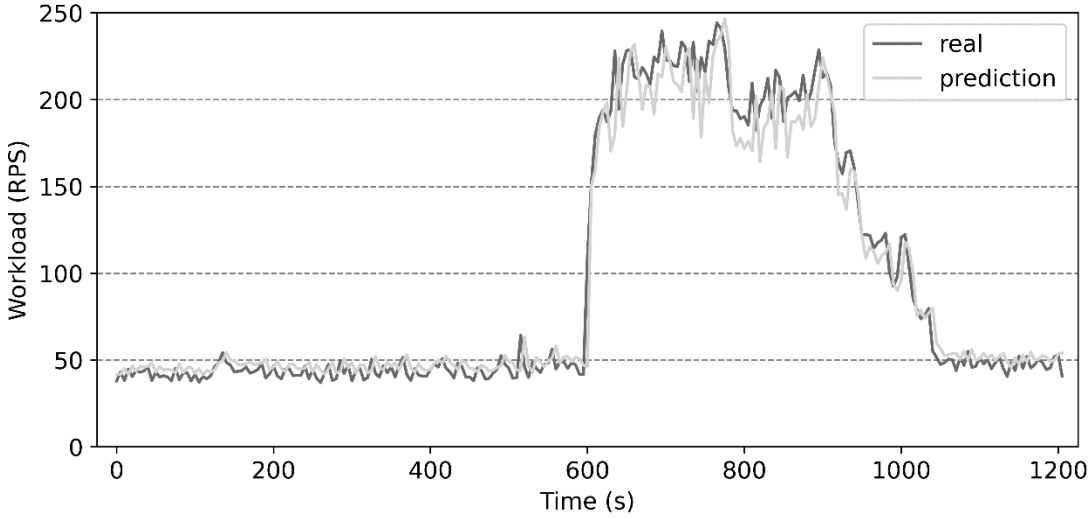


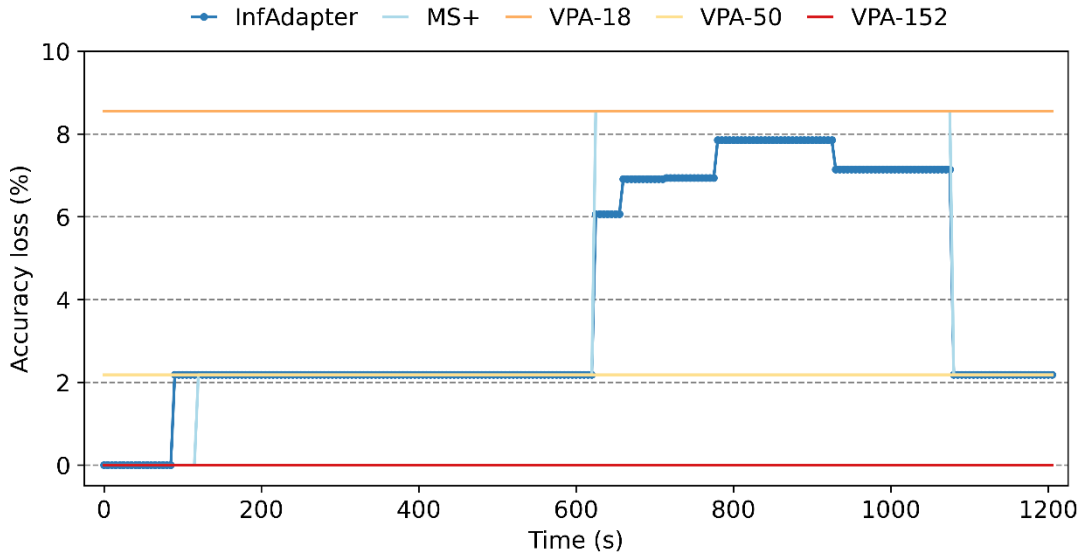
InfAdapter: P99-Latency evaluation



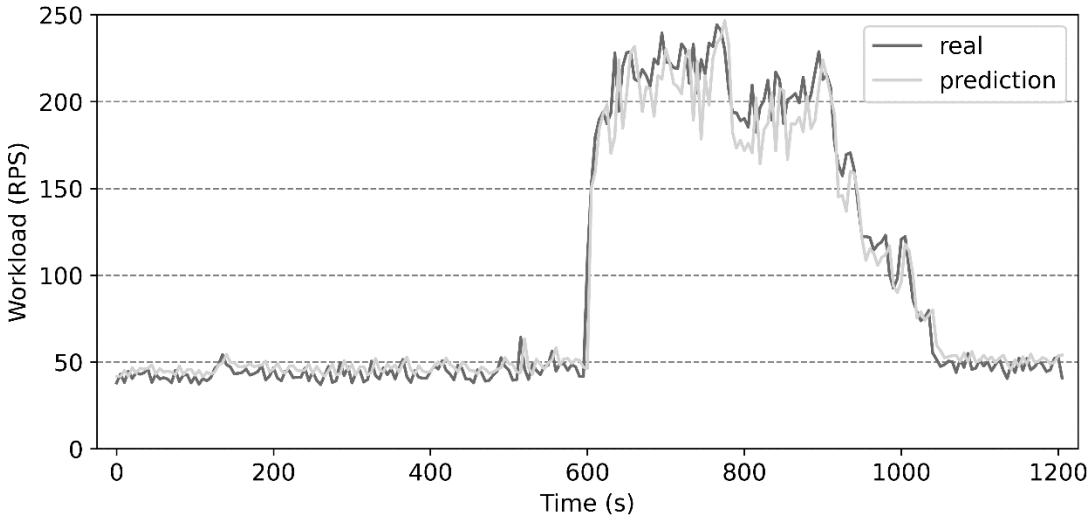


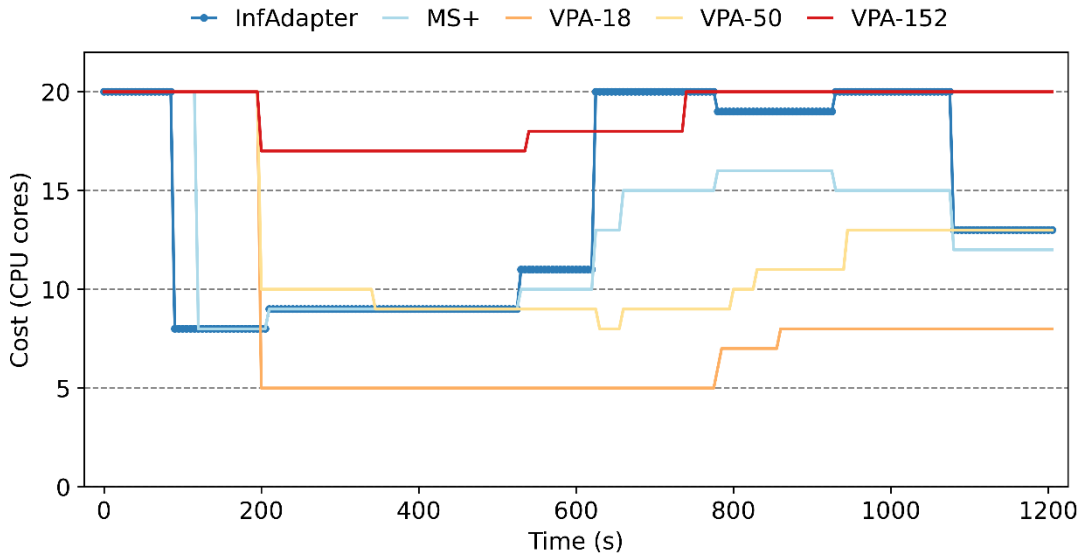
InfAdapter: Accuracy evaluation





InfAdapter: Cost evaluation





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

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

Highly Responsive! Cost-Efficient! Highly Accurate!

6

Takeaway

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

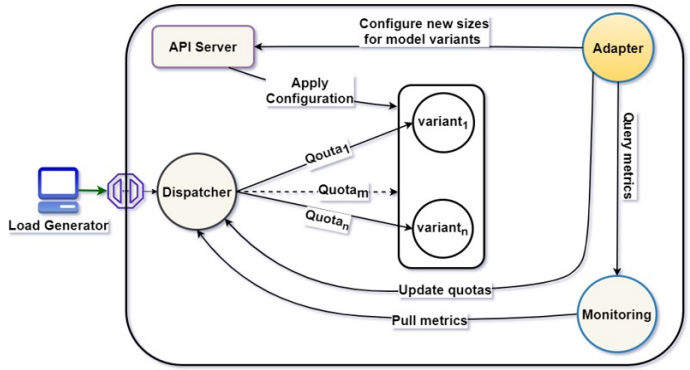
Model variants provide the opportunity to reduce resource costs while adapting to the dynamic workload.

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

InfAdapter!

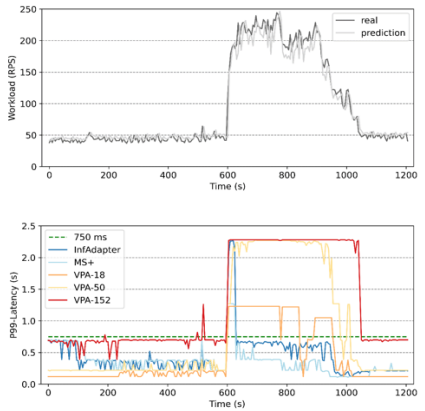
41

InfAdapter: Design



29

InfAdapter: P99-Latency evaluation



36

InfAdapter: Experimental evaluation

Compare aggregated metrics of latency SLO violation, accuracy and cost with other works on different β values to see how they perform on different accuracy-cost trade-off