

Adaptive Configuration Selection for Multi-Model Inference Pipelines in Edge Computing

Jinhao Sheng¹, Zhiqing Tang¹, Jianxiong Guo^{1,2}, and Tian Wang¹

¹Institute of Artificial Intelligence and Future Networks, Beijing Normal University, Zhuhai, China

²Guangdong Key Lab of AI & Multi-Modal Data Processing, BNU-HKBU United International College, Zhuhai, China
jsheng@cmu.edu.cn, {zhiqingtang, jianxiongguo, tianwang}@bnu.edu.cn

Abstract—The growing demand for real-time processing tasks is driving the need for multi-model inference pipelines on edge devices. However, cost-effectively deploying these pipelines while optimizing Quality of Service (QoS) and costs poses significant challenges. Existing solutions often neglect device resource constraints, focusing mainly on inference accuracy and cost efficiency. To address this, we develop a framework for configuring multi-model inference pipelines. Specifically: 1) We model the decision-making problem by considering the pipeline’s QoS, costs, and device resource limitations. 2) We create a feature extraction module using residual networks and a load prediction model based on Long Short-Term Memory (LSTM) to gather comprehensive node and pipeline status information. Then, we implement a Reinforcement Learning (RL) algorithm based on policy gradients for online configuration decisions. 3) Experiments conducted in a real Kubernetes cluster show that our approach significantly improve QoS while reducing costs and shorten decision-making time for complex pipelines compared to baseline algorithms.

Index Terms—Multi-model inference pipeline, Reinforcement learning, Edge computing, Kubernetes

I. INTRODUCTION

The rapid advancement of machine learning and large-scale deep learning models has led to a surge in demand for applications that can process tasks instantly. Applications such as Internet of Things (IoT) [1], autonomous driving [2], and healthcare [3] depend on complex model inference pipelines for stability and reliability in challenging situations. Many companies currently deploy their machine learning model inference pipelines on cloud platforms [4], but high latency and heavy reliance on network bandwidth often hinder real-time processing capabilities. Edge computing [5] addresses this issue by moving data processing, storage, and application services from centralized data centers to the network edge, significantly reducing latency and making it ideal for real-time processing scenarios [6] [7]. Consequently, deploying multi-model inference pipelines on edge devices is increasingly essential to meet these demands.

Deploying multi-model inference pipelines on edge devices can significantly reduce issues like latency, but finding a cost-effective way to implement these pipelines remains a major issue. Multi-model inference pipelines, which form a chain of machine learning models, raise several concerns

regarding Quality of Services (QoS) [8] and cost reduction. Unlike optimizations that focus on a single stage, optimizing an end-to-end inference pipeline requires a comprehensive approach that takes into account configuration adjustments across multiple steps. Previous works [9]–[11] have proposed solutions like efficient autoscaling, batching, and pipeline scheduling. These approaches address the aforementioned issues and also considered the dynamic nature of machine learning workloads. Existing methods often overlook critical factors such as inference accuracy and device resource constraints. Our work thoroughly considers these factors and provides a detailed model for configuring multi-model inference pipelines.

Configuring multi-model inference pipelines online to improve QoS and reduce costs holds great promise. However, several challenges need to be addressed. *The first challenge is how to effectively extract features from the dynamic load environment of these pipelines.* While existing methods like Fast, Accurate Autoscaling (FA2) [12] and Inference Pipeline Adaptation (IPA) system [13] offer some solutions, they often fail to comprehensively consider aspects like pipeline accuracy, cost control, and device resources. To tackle this issue, we propose a feature extraction module based on residual networks [14], which can deeply mine the state information of nodes and pipelines. Additionally, given the dynamic nature of workloads, accurately predicting incoming workloads is vital for adjusting the configurations of multi-model inference pipelines. To meet this need, we design a load prediction model based on Long Short-Term Memory (LSTM) [15]. The combination of the feature extraction module and the load prediction model provides robust real-time data support for configuration decisions.

Then, *the second challenge is how to effectively coordinate the decisions across multiple stages of the inference pipeline, including choices about replication factors, batch sizes, and model variations.* In the pipeline, the model selection at the earlier stages directly affects the best choices for the downstream models. The performance of the earlier models directly impacts the end-to-end latency and accuracy, which in turn limits the available options for downstream models. Although there are some heuristic algorithms like autoscaling [16] and model switching [17] trying to address this issue, these methods still fall short in coordinating decision-making for the various stages of the

inference pipeline. To tackle this problem, we propose an Online Pipeline Decision (OPD) algorithm based on policy gradients [18] that optimizes decision-making and flexibly adapts to changing workloads.

In this paper, we first model the online configuration selection problem for multi-model inference pipelines as a Markov Decision Process (MDP), aiming to maximize QoS while minimizing costs. Based on this, we propose a Reinforcement Learning (RL) algorithm using policy gradients for real-time configuration adjustments. To implement and evaluate this algorithm, we deploy multi-model inference pipelines in a real Kubernetes [19] cluster. Specifically, we use Selcon Core [20] and MLServer [21] to set up the pipeline, leveraging Prometheus [22] to monitor node status and track load information, which provides the necessary state data for the RL Agent’s decisions. We then adjust the configuration via the Kubernetes Python API to implement these decisions. The experimental results show that our algorithm outperforms existing baseline algorithms in enhancing QoS and reducing costs, while also significantly shortening decision-making time.

To sum up, the contributions of this paper are as follows:

- We model the online configuration decision problem in a multi-model inference pipeline as a MDP for the first time. Our goal is to maximize QoS while minimizing costs.
- We design a feature extraction module to extract information from nodes and the pipeline and an LSTM-based network to predict upcoming loads. We then present the OPD algorithm, enabling dynamic adaptation and optimization of decision-making.
- We validate the OPD algorithm’s effectiveness in a real Kubernetes cluster, demonstrating improved QoS, reduced costs, and significantly faster decision-making, surpassing all baseline algorithms.

II. RELATED WORK

Multi-stage Inference Serving: As the complexity and scale of machine learning applications grow, selecting configurations for model inference serving has garnered attention, especially in multi-stage inference services [9], [10], [12], [23]. Recent studies have investigated methods to improve inference performance in these systems. Crankshaw *et al.* [10] use a heuristic approach to optimize pipeline configurations based on Service Level Objectives (SLOs), effectively reducing end-to-end latency. Razavi *et al.* [12] automate scaling with target graph transformations and a dynamic programming algorithm. Ghafouri *et al.* [13] present IPA, a system that balances accuracy and cost objectives according to designer preferences. Salmani *et al.* [17] propose a strategy for dynamically switching between lightweight and heavy models based on workload, adapting to high-load conditions with models that have higher throughput but lower accuracy. Gunasekaran *et al.* [24] offer a cost-saving ensemble learning method that meets prior latency and accuracy requirements, despite being costly. However, existing

methods often overlook model switching, cost control, QoS, or resource constraints in edge devices.

RL in Edge Computing: In recent years, reinforcement learning (RL) algorithms have demonstrated outstanding performance in problems requiring long-term decision-making and strategy optimization. Many studies now apply RL methods to challenges in edge computing. Chen *et al.* [25] model data and service placement as MDP, using RL to achieve optimal placement decisions. Wang *et al.* [26] address service migration as observable MDP, implementing RL strategies for optimal migration. RL applications also extend to container migration [27], traffic control [6], task offloading [28], and container scheduling [29]. Building on this research, we analyze the feasibility of using RL for configuring a multi-model inference pipeline in edge computing.

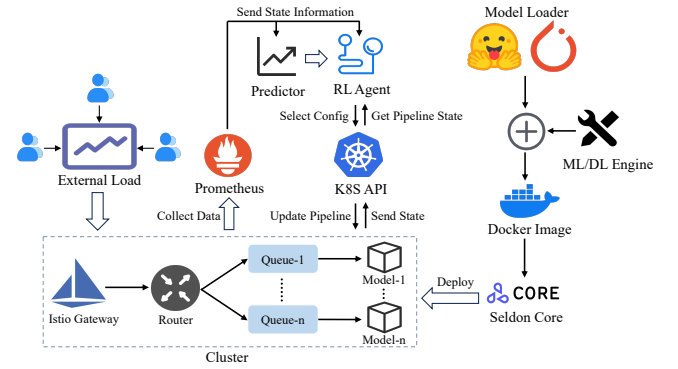


Fig. 1: System Overview.

III. SYSTEM DESIGN AND PROBLEM FORMULATION

A. System Design

This section provides an overview of the system’s main components, as illustrated in Fig. 1. The system consists of four core components: model loading, monitoring, pipeline system, and RL agent.

Model Loading: Users submit models for each pipeline stage, ensuring a balance of accuracy, latency, and resource usage. Model variants can be created using optimizers like TensorRT [30] and ONNX graph optimization [31], employing various quantization levels and Neural Architectural Search methods. Additionally, we store these model variants in an object storage service to minimize container creation time.

Monitoring: The monitoring daemon utilizes the highly available Prometheus [22] time-series database to periodically track the incoming load to the system and the pipeline.

Pipeline System: Once model loading is complete, the inference pipeline becomes fully operational, allowing users to send requests that traverse the models and return inference results. A centralized load balancer distributes these requests across the pipeline stages, each supported by a centralized queue to ensure predictable behavior and efficient latency modeling. GRPC [32] facilitates communication between the pipeline stages. Load balancing among containers within

the same stage is managed by Istio sidecar containers [33]. Each model container is deployed using Docker, sourced from a forked MLServer [21] and Seldon Core version, to implement gRPC web servers on Kubernetes. Detailed deployment steps are provided in Section V.

RL Agent: The RL agent periodically performs the following tasks: (1) retrieves incoming load and node information from the monitoring daemon, (2) uses a predictor to forecast the next reference load based on historical data, (3) obtains current pipeline configuration details via the Kubernetes Python API, and (4) selects and applies a configuration based on this information.

B. Definition of Metrics Over Pipeline

We examine a multi-model inference pipeline that includes key elements: task, accuracy, cost, and QoS [8]. These components define the system's operational and performance characteristics.

Task: The multi-model inference pipeline comprises a set of tasks denoted as $n \in \mathbf{N} = \{n_1, n_2, \dots, n_{|\mathbf{N}|}\}$, where $|\cdot|$ represents the number of tasks, such that $|\mathbf{N}|$ indicates the total number of tasks. Each task has an associated set of model variants $\mathbf{Z} = \{z_1, z_2, \dots, z_{|\mathbf{Z}|}\}$, which may differ in quantity. The accuracy and cost of task n using model variant z_i are denoted as $v_n(z_i)$ and $c_n(z_i)$, respectively. Additionally, the replication factor for task n is f_n , and the batch size is b_n . Throughput and latency for task n are represented by t_n and l_n , respectively.

Accuracy: To assess the accuracy of a multi-model inference pipeline, end-to-end accuracy is typically calculated using labeled data. However, this approach may not adequately capture the complexity of the semantic stages or the variability among models at each stage. To address this, we have adopted a heuristic approach inspired by [13], focusing on linear inference pipelines with a single input and output connected by a series of models. The accuracy of each model is pre-calculated offline, allowing us to determine the overall accuracy of the pipeline by:

$$V = \sum_{n \in \mathbf{N}, z_i \in \mathbf{Z}} v_n(z_i), \quad (1)$$

where $v_n(z_i)$ denotes the accuracy of the model variant z_i used in task n .

Cost: In our multi-model inference pipeline, cost refers to the CPU cores allocated to model variants. In Kubernetes, CPU resources are assigned by indicating the number of cores, which helps evaluate the pipeline's economic efficiency. The total cost of the pipeline can be calculated as:

$$C = \sum_{n \in \mathbf{N}, z_i \in \mathbf{Z}} f_n \times c_n(z_i), \quad (2)$$

where f_n denotes the number of replicas for task n , and $c_n(z_i)$ indicates the resource usage of model variant z_i for task n .

QoS: We have established specific QoS metrics for our multi-model inference pipelines, highlighting their unique attributes: accuracy, throughput, latency, and extra load as key performance indicators. We calculate QoS as follows:

$$Q = \begin{cases} \alpha \times V + \beta \times T - L - \gamma \times E, & \text{if } e \geq 0 \\ \alpha \times V + \beta \times T - L - \delta \times (-E), & \text{if } e < 0 \end{cases} \quad (3)$$

where α , β , γ , and δ are weighting parameters; T is the pipeline's throughput, defined as the minimum throughput across all tasks; L is the pipeline's latency, which is the sum of l_n ; E represents excess load, indicating the unprocessed portion of demand due to resource constraints. Positive E values signal unmet demand, while negative values indicate spare capacity. This definition of QoS enables us to assess the inference pipeline's performance from multiple dimensions.

C. Problem Formulation

Our objective is to improve QoS while simultaneously reducing costs. We aim to identify the optimal strategy that balances QoS and cost efficiency, defined as follows:

$$\begin{aligned} \max T &= Q - \lambda \times C, \\ \text{s.t.} \quad &\text{Eqs. (1), (2), (3),} \\ &0 < i \leq |\mathbf{Z}|, \forall z_i \in \mathbf{Z}, \\ &0 < f_n \leq F_{max}, \forall n \in \mathbf{N}, \\ &0 < b_n \leq B_{max}, \forall n \in \mathbf{N}, \\ &\sum_{n \in \mathbf{N}, z_i \in \mathbf{Z}} w_n(z_i) \times f_n \leq W_{max}, \end{aligned} \quad (4)$$

where F_{max} is the maximum allowed replication factor, B_{max} is the maximum permissible batch size in a pipeline, $w_n(z_i)$ represents the resource consumption of model variant z_i for task n , and W_{max} is the total resource capacity of the device.

Configuring multi-model inference pipelines is an NP-hard decision problem, meaning that finding or verifying an optimal solution is computationally difficult. Common approaches include heuristic algorithms, approximation techniques, and decomposition methods, each with specific limitations. RL algorithms present a promising alternative, particularly in complex decision environments where optimal strategies are not obvious. The configuration process in multi-model inference pipelines is dynamic, with each decision affecting current and future states. This sequential decision-making aligns well with MDP, making it suitable to model the problem as an MDP.

IV. OUR ALGORITHMS

A. Workload Prediction

We have developed an LSTM [15] model to predict the maximum workload for the next 20 seconds based on a time series of loads per second collected over the past 2 minutes. The model architecture includes a 25-unit LSTM layer followed by a one-unit dense output layer.

B. RL Settings

State: The state s_t comprises the node state and the pipeline state. The node state encompasses available resources, incoming loads, and predicted loads, while the

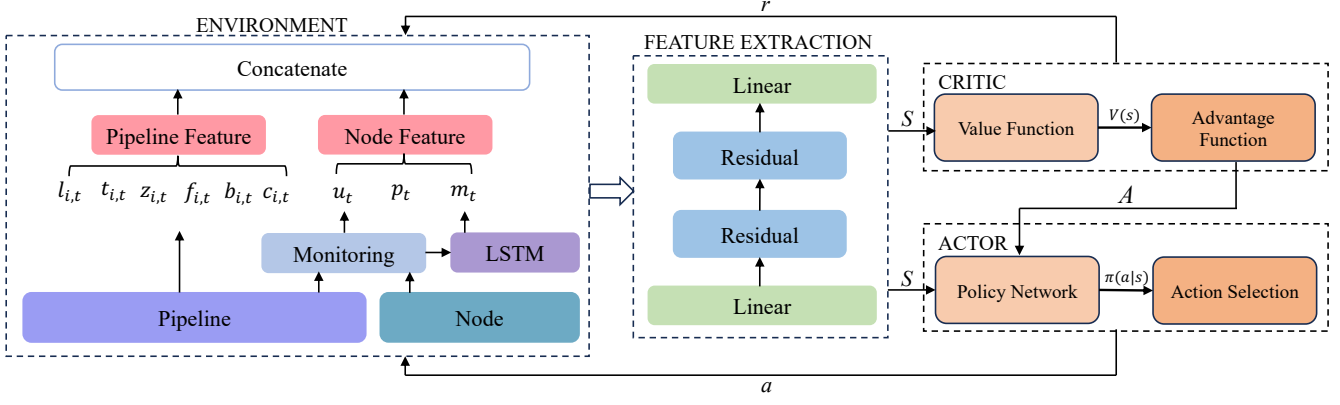


Fig. 2: Algorithm Overview.

pipeline state records each task's current model index, number of replicas, batch size, costs, latency, and throughput performance. Thus, the state s_t is defined as:

$$s_t = [(u_t, p_t, m_t, l_{1,t}, t_{1,t}, z_{1,t}, f_{1,t}, b_{1,t}, c_{1,t}), \dots, (u_t, p_t, m_t, l_{n,t}, t_{n,t}, z_{n,t}, f_{n,t}, b_{n,t}, c_{n,t})]. \quad (5)$$

Action: The action a_t encompasses the configuration for each pipeline task, including the model variant index, number of replicas, and batch size. Thus, a_t can be defined as:

$$a_t = [(z_{1,t}, f_{1,t}, b_{1,t}), \dots, (z_{n,t}, f_{n,t}, b_{n,t})]. \quad (6)$$

When the task changes, the action space must be modified to match the configuration options of each pipeline task.

Reward: Given that various workloads require distinct configurations, we have optimized the reward function to improve the stability and effectiveness of policy gradient algorithms while achieving our desired objectives. The reward function is defined as follows:

$$r_t = Q - \beta \times C - \gamma \times B, \quad (7)$$

where the parameter β represents the weight assigned to the cost, γ is the penalty coefficient for the batch sizes, and B denotes the maximum batch sizes b_n for each task in a pipeline. γ is employed to prevent excessively large batch sizes, ensuring that configurations enhance QoS while maintaining a reasonable scale.

C. Online Pipeline Decision

In this section, we introduce the Online Pipeline Decision (OPD) algorithm, detailing the feature extraction process and the training methodology.

Overview: The OPD algorithm framework is shown in Fig. 2. The system retrieves the resource status, load conditions, and task configurations from the environment. This information is fed into feature extraction and processed by the policy network to make decisions and calculate decision-time, as detailed in Algorithm 1. Rewards are generated based on the actions taken, and the policy network alongside the value functions is updated using policy gradient algorithms [18].

Algorithm 1 The OPD Algorithm

Input: L, P, M

Output: H

```

Deploy Pipeline in Cluster;
Load the OPD Agent;
Reset environment and get state  $s_0$ ;
Incoming workloads;
for  $t = 1, 2, \dots$  do
    Predict the coming workload;
    Get the state  $s_t$  by Eq. (5);
    Select  $a_t$  according to the OPD agent;
    Calculate the decision time  $d_t$ ;
    Apply the action  $a_t$  and convert it into a configuration
    that can be applied to the cluster;
    Apply new configuration by Kubernetes Python API;
    Calculate reward  $r_t$ ;
    Get the next state  $s_{t+1}$ ;
end for
Compute the cumulative decision time:  $H = \sum d_t$ ;

```

Feature Extraction: The feature consists of two main components: node features and pipeline features, which are combined into a unified feature vector. Raw data from node and pipeline states undergoes processing through a fully connected layer to reduce dimensionality to a manageable intermediate form. This processed data is then refined through several residual blocks [14].

Training: The OPD algorithm focuses on policy optimization, where a policy, typically denoted as π , guides the agent's action choices, expressed as $a(t) \sim \pi(\cdot|s(t))$. In multi-model inference scenarios, the probability of the decision process is defined as:

$$P(\gamma|\pi) = \rho_0(s_0) \prod_{t=0}^{T-1} P(s_{t+1}|s_t, a_t) \pi(a_t|s_t), \quad (8)$$

where $\rho_0(s_0)$ is the start-state distribution.

In policy gradient algorithms, the most commonly used gradient estimator \hat{g} has the form of:

$$\hat{g} = \hat{\mathbb{E}}_t[\nabla_{\theta} \log \pi_{\theta}(a_t|s_t) \hat{A}_t], \quad (9)$$

Algorithm 2 Training of the OPD Algorithm

Input: P, L **Output:** a_t

```
Initialize policy network  $\pi_\theta$ ;
Initialize replay memory  $\mathbf{D} = \emptyset$ ;
Initialize expert optimizer as expert_model;
Define expert_frequency as  $f$ ;
for  $e \leftarrow 1, 2, \dots$  do
  Reset the environment and obtain the initial state  $s_0$ ;
  while not completed processing of  $L$  do
    if  $e \% f == 0$ 
       $a_t \leftarrow$  action from expert_model given  $s_t$ ;
    else
       $a_t \leftarrow$  action from policy  $\pi_\theta$  given  $s_t$ ;
    end if
    Execute action  $a_t$  in the environment;
    Observe new state  $s_{t+1}$  and calculate reward  $r_t$ ;
    Store  $\{s_t, a_t, r_t\}$  in  $\mathbf{D}$  and compute target value
     $\hat{V}_\pi(s_t)$ ;
  end while
  for  $t = 1, 2, \dots, T$  do
    Compute  $\hat{A}_1, \dots, \hat{A}_T$ ;
    Compute loss  $L(\theta)$  by Eq. (11);
    Optimize the network by mini-batch SGD with
    Adam optimizer;
  end for
  Update weights from time to time  $\pi_{old} \leftarrow \pi_\theta$ ;
end for
```

which is obtained by differentiating the loss function $L^{PG}(\theta)$ of policy gradient:

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]. \quad (10)$$

However, traditional policy gradient methods can be inefficient and lead to disruptive updates, resulting in instability. To mitigate these issues, Proximal Policy Optimization (PPO) [34] uses a clipped surrogate objective:

$$L_t(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)], \quad (11)$$

where c_1, c_2 are coefficients, and S denotes an entropy bonus, and L_t^{VF} is a squared-error loss. The $L_t^{CLIP}(\theta)$ is denoted as:

$$L_t^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \quad (12)$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$ is the policy ratio, \hat{A}_t is the estimated advantage function, and ϵ is a small constant that limits the range of $r_t(\theta)$ to prevent training instabilities from large updates.

To tackle the issues of sparse training data and slow initial convergence, we have integrated expert guidance into our training. The training process for the OPD algorithm is detailed in Algorithm 2.

D. Computational Complexity Analysis

The OPD algorithm consists of four components: state observation, action selection, reward calculation, and net-

work update. Below is an analysis of the computational complexity for each part. First, the state is defined in Eq. (5), with a complexity of $O(|N|)$. Second, the complexity of action selection for all pipeline tasks is also $O(|N|)$. The complexity of reward calculation remains constant regardless of the number of tasks, resulting in $O(1)$. Finally, feature extraction occurs through fully connected layers within residual blocks, where each layer spans L layers and contains G neurons. The complexity for this section is $O(|N| \times G + L \times G^2)$.

The total experience used for each network update is constant, while the ratio of expert to agent experience varies. Other environmental operations have little effect on the computational complexity. Thus, the complexity of the OPD algorithm is expressed as $O(|N| + |N| \times G + L \times G^2)$.

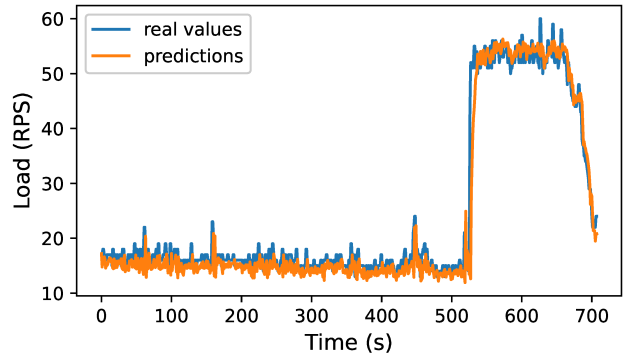


Fig. 3: Lstm Prediction.

V. SYSTEM IMPLEMENTATION

A multi-model inference pipeline is deployed on a Kubernetes cluster, a popular open-source platform for orchestrating containerized applications.

A. Implementation Issues

Network Communication: The system processes inference requests sequentially across multiple models, which can introduce latency. To mitigate this, it uses gRPC [32] for inter-stage communication, optimizing data transfer efficiency and reducing delays.

Multi-Model Dependency Management: Different models may need various library versions or runtime environments, leading to dependency conflicts. Deploying models in Docker containers isolates these environments, preventing version conflicts and ensuring stable system operation.

Resource Management and Scheduling: A critical challenge during high-load scenarios is ensuring efficient resource utilization among multiple containers at the same stage. We implement Istio sidecar containers [33] for load balancing and employ our OPD algorithm to dynamically adjust pipeline configurations based on available resources, maximizing resource usage and preventing any instance from being resource-constrained.

B. Deployment of Multi-Model Inference Pipelines

Machine learning models can be encapsulated using MLServer and Seldon Core tools to create Docker images. These images are deployed on a Kubernetes cluster via a “SeldonDeployment” YAML file, which outlines deployment details such as resource requirements, replica counts, exposed ports, and other relevant information.

In an Istio-supported environment, the multi-model inference pipeline is managed through Istio’s routing rules, utilizing “Virtual Service” and “Destination Rule” to direct and optimize data flow between models, thereby enhancing the microservices’ security, reliability, and efficiency.

Once deployed, Seldon Core’s management interface and APIs allow for monitoring, management, and optimization of the model’s operational status. For detailed performance monitoring and log analysis, integration with Prometheus and Grafana is also possible.

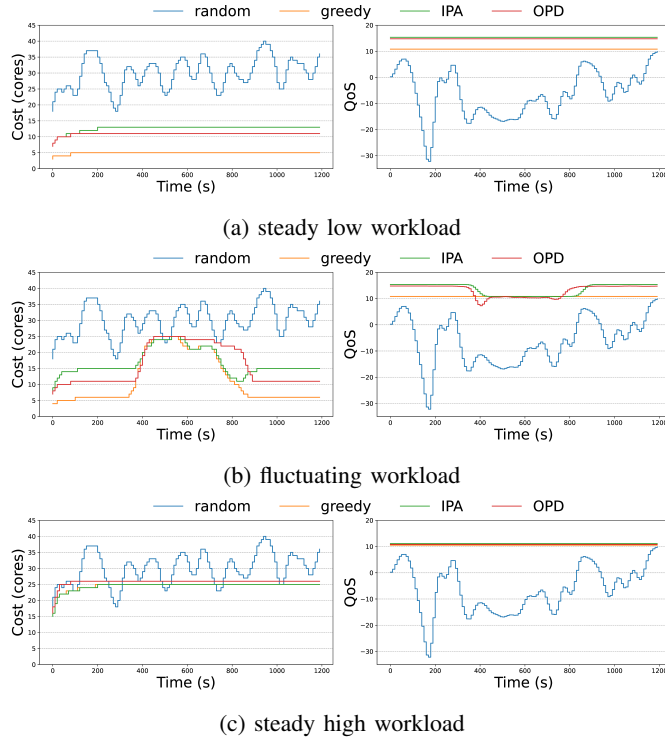


Fig. 4: Temporal analysis under different workloads.

VI. PERFORMANCE EVALUATION

A. Experimental Settings

Hardware: We evaluate the OPD algorithm using three physical machines, each equipped with an Intel i9-10900K 10-core CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 2070 Super graphics card.

Predictor: The LSTM module has been trained to predict workloads in under 50 milliseconds. As shown in Fig. 3, the model demonstrates high accuracy in testing, achieving a Symmetric Mean Absolute Percentage Error (SMAPE) [14] of only 6%, comparable to predictors in similar systems.

Pipelines: Previous studies frequently relied on customized pipelines, which posed challenges in production settings. In contrast, our pipelines prioritize flexibility and versatility, incorporating insights from earlier research and industrial applications, as outlined in [13]. Each task within our pipeline is supported by a selection of models tailored to specific requirements.

Baselines: Several baselines are conducted for performance comparison, detailed as follows.

- 1) **Random.** The algorithm randomly selects configurations for each pipeline task, promoting diversity and exploration in selection.
- 2) **Greedy.** A greedy algorithm chooses the configuration for each pipeline task to minimize costs while adhering to available resource constraints.
- 3) **IPA [13].** This algorithm uses a solver to select configurations for each task in a pipeline but ignores resource constraints. We have enhanced the algorithm to factor in resource availability during configuration selection.

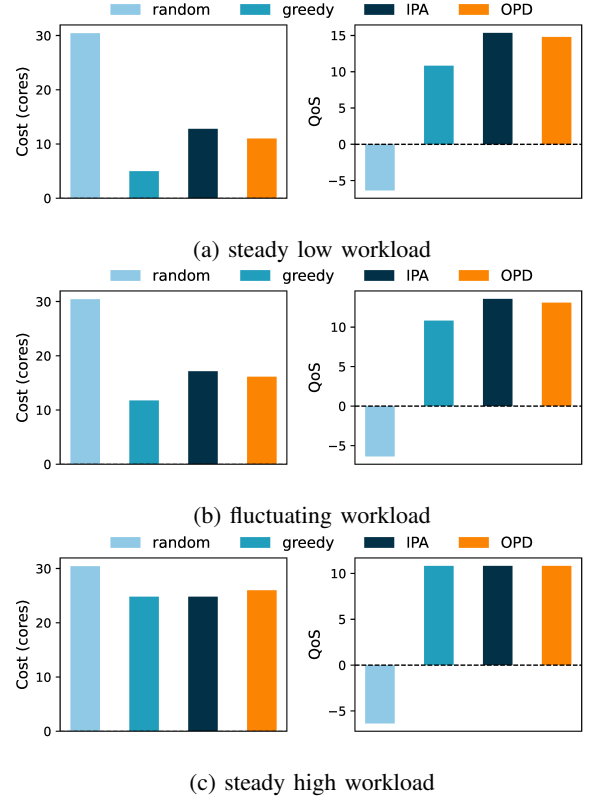


Fig. 5: Performance analysis under different workloads.

B. Experimental Results

We evaluate the performance of the proposed OPD algorithm through a series of experiments in the Kubernetes cluster, comparing it with baseline algorithms.

Performance with different workloads: To minimize overhead from frequent switching, we established a 10-second adaptation interval and meticulously recorded performance data during each 1200-second operational cycle

to evaluate four algorithms under various workloads. To ensure reproducibility and consistency, we fix the seed for all random generators.

Steady low load: Fig. 4(a) shows the performance of various algorithms under steady low load, offering a detailed comparison. The analysis indicates that all algorithms, except the random one, maintain consistent costs and QoS in this environment. While the greedy algorithm has the lowest cost, its QoS is relatively poor. In contrast, the IPA algorithm, despite being the most expensive, delivers the highest QoS. Our OPD algorithm strikes a balance between cost and QoS, positioning itself between greedy and IPA, and nearly matching IPA in service performance.

Fluctuating load: Fig. 4(b) shows the algorithms' performance under fluctuating loads. Like in stable low-load conditions, the greedy algorithm excels in cost but fails to manage QoS during load fluctuations. The IPA algorithm offers stable services but is limited by its high cost in cost-sensitive settings. In contrast, the OPD algorithm effectively balances QoS and cost under varying loads, showcasing its adaptability.

Steady high load: Fig. 4(c) illustrates performance under steady high load conditions. In these environments, the high volume of task requests leads to increased costs for all algorithms. The random algorithm demonstrates significant cost fluctuations, reflecting instability in its resource allocation. In contrast, the greedy algorithm, IPA, and OPD maintain similar cost control and nearly identical QoS, suggesting they effectively balance cost and QoS to meet task demands.

Figs. 5(a), 5(b), and 5(c) show the average cost and QoS of various algorithms across a single workload cycle. Under steady low-load conditions, the OPD algorithm has a cost 120% higher than the greedy algorithm but improves QoS by 36%. Compared to the IPA method, it reduces costs by 16% with only a minor 3.8% drop in QoS. In fluctuating load conditions, OPD's cost is 37% higher than the greedy algorithm's, yet it enhances QoS by 21%. Relative to IPA, it decreases costs by 6% while slightly lowering QoS by 3%. In stable high-load scenarios, the greedy, IPA, and OPD algorithms exhibit nearly identical costs and QoS due to the high volume of requests, demonstrating OPD's effectiveness in balancing costs and QoS across different load conditions.

In short, the random algorithm exhibits significant cost and QoS fluctuations under different loads, indicating instability. The greedy algorithm is cost-effective but requires improved QoS. In contrast, both the IPA and OPD algorithms offer stable QoS. However, the IPA overlooks resource constraints and requires quicker decision-making. The OPD algorithm addresses these issues, optimally balancing cost and QoS while outperforming all baseline algorithms.

Performance with different pipelines: To evaluate decision-making times across pipelines of varying complexities, we analyze four distinct pipelines with differing stages and model variants. Given the limited value of comparisons with random and greedy algorithms, we focus on the IPA and OPD methods. As shown in Fig. 6, IPA's decision-making time increases with pipeline complexity,

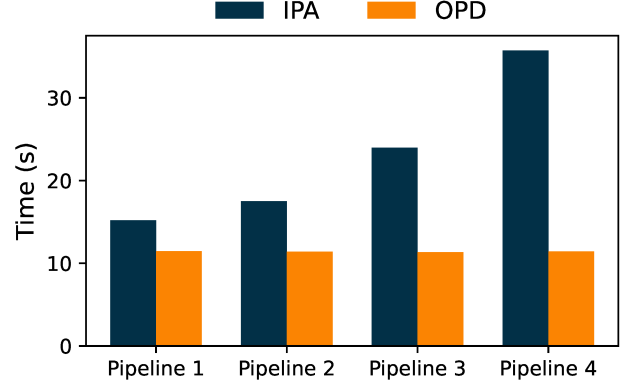


Fig. 6: Different pipelines decision time.

while OPD's remains steady. As the number of tasks and model variants per stage increases, OPD's decision-making efficiency significantly surpasses IPA, with improvements of 32.5%, 53.5%, 111.6%, and 212.8% in processing a single workload cycle. Thus, the OPD algorithm effectively reduces decision-making time in complex multi-model inference pipelines, demonstrating the efficient applicability of the algorithm in large-scale and complex scenarios.

Security of the OPD algorithm: In a distributed environment, dynamically adjusting the configurations of a multi-model inference pipeline may introduce potential security risks. To prevent these risks, we set restrictions such as the number of replicas and batch sizes in the OPD algorithm to prevent malicious users or attackers from manipulating the algorithm to select inappropriate configurations, thereby avoiding problems such as system overload and service interruption.

Convergence of the OPD algorithm: To illustrate the convergence of the OPD algorithm, Fig. 7 displays the training loss, value loss, and mean episode reward. As training progresses, both the training loss and value function loss decrease rapidly, eventually stabilizing around a specific value, indicating convergence. The figure also shows that the OPD algorithm's reward converges to a higher value. Overall, the OPD algorithm exhibits a remarkably rapid convergence rate, demonstrating its ability to learn optimal policies in a short timeframe.

VII. CONCLUSION

We developed an online configuration decision algorithm for multi-model inference pipelines. First, we thoroughly modeled the OPD problem, taking into account cost, QoS, and device resource limitations. Next, we introduced a feature extraction method using a residual network to effectively capture node and pipeline state features. Finally, we proposed an OPD algorithm with an LSTM predictor based on policy gradient reinforcement learning for making configuration decisions. Experiments conducted on a real Kubernetes cluster showed that our algorithm outperforms baseline methods. Future work will focus on optimizing

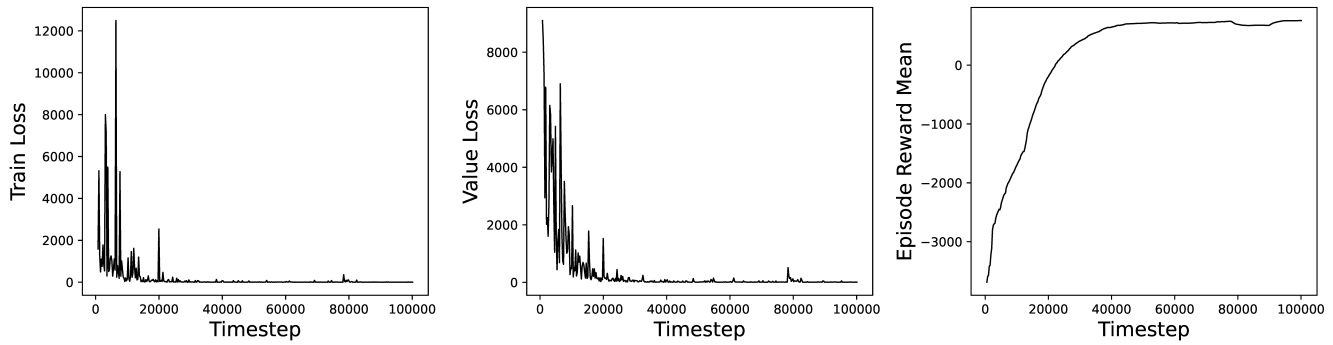


Fig. 7: Training Loss, Value Loss, and Reward of the OPD algorithm.

GPU resource allocation for multi-model inference pipelines in the Kubernetes cluster to enhance inference speed, reduce latency, and achieve higher QoS.

REFERENCES

- [1] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of things (iot): A literature review," *Journal of Computer and Communications*, vol. 3, no. 5, pp. 164–173, 2015.
- [2] J. E. Naranjo, C. González, R. García, T. de Pedro, and R. E. Haber, "Power-steering control architecture for automatic driving," *Ieee transactions on intelligent transportation systems*, vol. 6, no. 4, pp. 406–415, 2005.
- [3] M. Wehde, "Healthcare 4.0," *IEEE Engineering Management Review*, vol. 47, no. 3, pp. 24–28, 2019.
- [4] C. Gong, J. Liu, Q. Zhang, H. Chen, and Z. Gong, "The characteristics of cloud computing," in *2010 39th International Conference on Parallel Processing Workshops*, pp. 275–279, IEEE, 2010.
- [5] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [6] Z. Ning, K. Zhang, X. Wang, M. S. Obaidat, L. Guo, X. Hu, B. Hu, Y. Guo, B. Sadoun, and R. Y. Kwok, "Joint computing and caching in 5g-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 5201–5212, 2020.
- [7] Z. Tang, F. Mou, J. Lou, W. Jia, Y. Wu, and W. Zhao, "Joint resource overbooking and container scheduling in edge computing," *IEEE Transactions on Mobile Computing*, 2024.
- [8] A. Campbell, G. Coulson, and D. Hutchison, "A quality of service architecture," *ACM SIGCOMM Computer Communication Review*, vol. 24, no. 2, pp. 6–27, 1994.
- [9] Y. Hu, W. Pang, X. Liu, R. Ghosh, B. Ko, W.-H. Lee, and R. Govindan, "Rim: Offloading inference to the edge," in *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pp. 80–92, 2021.
- [10] D. Crankshaw, G.-E. Sela, X. Mo, C. Zumar, I. Stoica, J. Gonzalez, and A. Tumanov, "Inferline: latency-aware provisioning and scaling for prediction serving pipelines," in *Proceedings of the 11th ACM Symposium on Cloud Computing*, pp. 477–491, 2020.
- [11] R. S. Kannan, L. Subramanian, A. Raju, J. Ahn, J. Mars, and L. Tang, "Grand slam: Guaranteeing slas for jobs in microservices execution frameworks," in *Proceedings of the Fourteenth EuroSys Conference 2019*, pp. 1–16, 2019.
- [12] K. Razavi, M. Luthra, B. Koldehofe, M. Mühlhäuser, and L. Wang, "Fa2: Fast, accurate autoscaling for serving deep learning inference with sla guarantees," in *2022 IEEE 28th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 146–159, IEEE, 2022.
- [13] S. Ghafouri, K. Razavi, M. Salmani, A. Sanaee, T. L. Botran, L. Wang, J. Doyle, and P. Jamshidi, "[solution] ipa: Inference pipeline adaptation to achieve high accuracy and cost-efficiency," *Journal of Systems Research*, vol. 4, no. 1, 2024.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [15] S. Hochreiter, "Long short-term memory," *Neural Computation MIT Press*, 1997.
- [16] K. Rzdca, P. Findeisen, J. Swiderski, P. Zych, P. Broniek, J. Kusmerek, P. Nowak, B. Strack, P. Witusowski, S. Hand, *et al.*, "Autopilot: workload autoscaling at google," in *Proceedings of the Fifteenth European Conference on Computer Systems*, pp. 1–16, 2020.
- [17] M. Salmani, S. Ghafouri, A. Sanaee, K. Razavi, M. Mühlhäuser, J. Doyle, P. Jamshidi, and M. Sharifi, "Reconciling high accuracy, cost-efficiency, and low latency of inference serving systems," in *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pp. 78–86, 2023.
- [18] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.
- [19] Kubernetes. [Online]. Available: <https://kubernetes.io/docs/home/>.
- [20] Seldon Core. Available: <https://github.com/SeldonIO/seldon-core>.
- [21] MLServer. Available: <https://github.com/SeldonIO/MLServer>.
- [22] Prometheus. [Online]. Available: <https://prometheus.io>.
- [23] Z. Zhou, Y. Zhang, and C. Delimitrou, "Aquate: Qos-and-uncertainty-aware resource management for multi-stage serverless workflows," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 1–14, 2022.
- [24] J. R. Gunasekaran, C. S. Mishra, P. Thinakaran, B. Sharma, M. T. Kandemir, and C. R. Das, "Cocktail: A multidimensional optimization for model serving in cloud," in *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pp. 1041–1057, 2022.
- [25] L. Chen, J. Xu, S. Ren, and P. Zhou, "Spatio-temporal edge service placement: A bandit learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8388–8401, 2018.
- [26] J. Wang, J. Hu, G. Min, Q. Ni, and T. El-Ghazawi, "Online service migration in mobile edge with incomplete system information: A deep recurrent actor-critic learning approach," *IEEE Transactions on Mobile Computing*, vol. 22, no. 11, pp. 6663–6675, 2022.
- [27] Z. Tang, F. Mou, J. Lou, W. Jia, Y. Wu, and W. Zhao, "Multi-user layer-aware online container migration in edge-assisted vehicular networks," *IEEE/ACM Transactions on Networking*, 2023.
- [28] J. Wang, J. Hu, G. Min, W. Zhan, A. Y. Zomaya, and N. Georgalas, "Dependent task offloading for edge computing based on deep reinforcement learning," *IEEE Transactions on Computers*, vol. 71, no. 10, pp. 2449–2461, 2021.
- [29] H. Cui, Z. Tang, J. Lou, W. Jia, and W. Zhao, "Latency-aware container scheduling in edge cluster upgrades: A deep reinforcement learning approach," *IEEE Transactions on Services Computing*, 2024.
- [30] NVIDIA TensorRT. [Online]. Available: <https://developer.nvidia.com/tensorrt>.
- [31] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*, pp. 291–326, Chapman and Hall/CRC, 2022.
- [32] GRPC. [Online]. Available: <https://grpc.io/>.
- [33] Istio. [Online]. Available: <https://istio.io/>.
- [34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.