Noname manuscript No.
(will be inserted by the editor)

On the Impact of Black-box Deployment Strategies for Edge AI on Latency and Model Performance

Empirical Study on ONNX

Jaskirat Singh · Emad Fallahzadeh · Bram Adams · Ahmed E. Hassan

Received: date / Accepted: date

Context: Cloud-based black-box model deployment faces challenges related to latency and privacy due to data transmission across Wide Area Networks. On the other hand, Mobile-based black-box deployment prioritizes privacy at the expense of higher latency due to limited computational resources. To address these issues, Edge AI enables the deployment of black-box models across Mobile, Edge, and Cloud devices using a wide range of operators able to distribute a model's components, terminate inference early, or even quantize a model's computations, offering latency and privacy benefits. Existing surveys classify Edge AI model inference techniques into eight families, including Quantization, Early Exiting, and Partitioning, but they often treat these operators in isolation, overlooking their potential synergies and practical integration in real-world scenarios. Deciding what combination of operators to use across the Edge AI tiers to achieve specific latency and model performance requirements is still an open question for MLOps Engineers. *Objective:* This study aims to empirically assess the accuracy vs inference time trade-off of different black-box Edge AI deployment strategies, i.e., combinations of deployment operators and deployment tiers. *Method:* In this paper, we conduct inference experiments involving three deployment operators (i.e., Partitioning, Quantization, Early Exit), three deployment tiers (i.e., Mobile, Edge, Cloud) and their combinations on four widely-used Computer-Vision models to investigate the optimal strategies from the point of view of MLOps developers. The analysis is conducted in a containerized environment using CUDA for Cloud GPU acceleration and ONNX for model interoperability, covering a wide range of network bandwidths. *Results:* Our findings suggest that Edge deployment using the hybrid Quantization + Early Exit operator could be preferred over Non-Hybrid operators (Quantization/Early Exit on Edge, Partition on Mobile-Edge) when faster latency is a concern at medium accuracy loss. However, when minimizing accuracy loss is a concern, MLOps Engineers should prefer using only a Quantization operator on Edge at a latency reduction or increase, respectively over the Early Exit/Partition (on Edge/Mobile-Edge) and Quantized Early Exit (on Edge) operators. In scenarios constrained by Mobile CPU/RAM resources, a preference for Partitioning across Mobile and Edge tiers is observed over Mobile deployment. For models with smaller input data samples (such as FCN), a network-constrained Cloud deployment can

Jaskirat Singh
School of Computing, Queen's University, Kingston, Ontario, Canada
E-mail: jaskirat.s@queensu.ca
https://orcid.org/0009-0001-0112-7674

Emad Fallahzadeh
School of Computing, Queen's University, Kingston, Ontario, Canada.
E-mail: emad.fallahzadeh@queensu.ca
https://orcid.org/0009-0005-5024-4868

Bram Adams
School of Computing, Queen's University, Kingston, Ontario, Canada.
E-mail: bram.adams@queensu.ca
https://orcid.org/0000-0001-7213-4006

Ahmed E. Hassan
School of Computing, Queen's University, Kingston, Ontario, Canada
E-mail: hassan@queensu.ca
https://orcid.org/0000-0001-7749-5513

also be a better alternative than Mobile/Edge deployment and Partitioning strategies. For models with large input data samples (ResNet, ResNext, DUC), an Edge tier having higher network/computational capabilities than the Cloud/Mobile tier can be a more viable option than Partitioning and Mobile/Cloud deployment strategies. Smaller input data-sized models like FCN fit well in the Cloud, even with low bandwidth ($\leq$10 Mbps). Larger input data-sized models, like ResNe(x)t and DUC, need more bandwidth ($\geq$50 Mbps) for Cloud latency convergence. Partitioned-based strategies for large intermediate-sized models like FCN and DUC also need at least 50 Mbps for latency convergence. Overall, the Cloud tier performs better than the Edge and Mobile tiers for Non-Partitioning operators when the MEC bandwidth is at least 50 Mbps. However, its latency performance declines in lower bandwidth scenarios. Furthermore, Mobile-Edge Partitioning-based strategies are a better alternative compared to Mobile-Cloud and Edge-Cloud alternatives.

Keywords Edge AI · Deployment Strategies · Inference Latency · Model Performance

1 Introduction

Artificial Intelligence (AI) on the Edge (also "Edge Intelligence" or "Edge AI") [129], an interdisciplinary field derived from Edge computing and AI, receives a tremendous amount of interest from both the industry and academia. This is primarily due to its low latency, privacy preservation, and potential independence from network connectivity. Edge AI leverages widespread Edge resources instead of relying solely on Cloud or Mobile, leading to more efficient AI insights for inference and training tasks. For example, in our experiments, we consider inference tasks in a typical Edge AI environment involving an Edge device (tier) near a resource-scarce Mobile device (tier) and a resource-abundant Cloud device (tier) far from the Edge device (tier).

Traditional monolithic deployments such as deploying large AI models entirely on a Cloud or a Mobile tier affect the overall performance in terms of Key Performance Indicators (KPIs). For example, deploying entire AI models on the Cloud provides faster computation in model inference due to the available GPU resources. However, it leads to high transmission latency, monetary cost, and privacy leakage when transmitting large amounts of input data across the Wide-Area Network (WAN) to a centralized data center for AI applications (e.g., real-time video analytics). On-device inference, running entire AI applications on the Mobile tier to process the input data locally, provides data privacy protection but suffers from high computation latency because many AI applications require high computational power that significantly outweighs the capacity of resource-constrained Mobile tiers [104].

Edge computing essentially pushes Cloud-like services to network Edge servers that are in closer proximity to Mobile tiers and data sources [110]. This offers several benefits compared to the traditional Cloud-based paradigm (i.e., low transmission latency, data privacy protection, and low monetary cost) and the Mobile-based paradigm (i.e., faster computational latency). However, this comes at the expense of increased computational latency compared to the Cloud and a higher data privacy threat compared to the Mobile.

Various operators for Edge AI model inference are proposed to address the above challenges faced by monolithic Mobile, Edge, and Cloud deployments. Zhou et al. [129] provide a detailed survey on seven major families of deployment operators. Among the three model optimization operator families, Model Compression includes operators like Weight Pruning, Knowledge Distillation, and Quantization to reduce computation and storage; Model Partition provides computational offloading across the tiers and latency/energy-oriented optimization; and Model Early-Exit performs partial DL model inference at Early Exit points, trading accuracy for speed. Among the other families, Edge Caching focuses on reusing previous results of the same inference task for faster response, Input Filtering detects differences between inputs to avoid redundant computation, Multi-Tenancy supports scheduling multiple DL-based tasks in a resource-efficient manner, and Model Selection uses Input-oriented, accuracy-aware optimization.

Currently, MLOps Engineers continuously experiment with different combinations of operators to find an optimal balance in latency and model prediction performance. White-box operators require substantial time for re-training or fine-tuning a model by changing weights and structure. This demands a deep understanding of the internal workings of the model, including its architecture, parameters, and training process. Furthermore, the resulting model may behave unpredictably compared to previously tested versions. In contrast, black-box operators allow quicker adaptation to models without requiring an in-depth understanding of their internal architecture or parameters. These operators apply transformations to pre-trained models and are often favored in scenarios where model transparency is limited, especially in DNN models.

Given black-box Edge AI operators, the challenge becomes: 1) Where (on which tiers) to deploy models in an Edge AI setting; and 2) How to post-process models using operators to make them compatible with those tiers. The combination of a choice of tier and operator forms an Edge AI deployment strategy. While there are many deployment operators and tiers, MLOps Engineers currently rely on trial and error to find the best configuration. Hence, this study initiates a catalog of empirical data to eventually enable recommendation systems that assist MLOps Engineers in deciding the most appropriate deployment strategy for their context.

The main contribution of this study is an in-depth empirical comparison between competing Edge AI deployment strategies to suggest recommendations for deploying DNN models for MLOps Engineers. Specifically, we compare strategies mapping common black-box deployment operators, including the Identity operator (which serves as a baseline with no model transformation), Partitioned, Early Exit, Quantized, and their combinations (Quantized Early Exit [QE] and Quantized Early Exit Partitioned [QEP]) to three common deployment tiers (i.e., Mobile, Edge, and Cloud) and their combinations in an Edge AI environment. Second, for each of the Edge AI deployment strategies, we evaluate the end-to-end (round-trip) latency in an Edge AI setup (Mobile, Edge, and Cloud tiers). Third, we focus on measuring the latency of deployment strategies across a wide range of varying input (i.e., image) sizes using sequential inference requests. Our study analyzes the optimal trade-off in terms of inference latency and accuracy among competing Edge AI deployment strategies. We address the following research questions:

- RQ1: What is the impact of monolithic deployment in terms of inference latency and accuracy across the considered tiers?
- RQ2: What is the impact of the Quantized operator in terms of inference latency and accuracy within and across the considered tiers?
- RQ3: What is the impact of the Early Exit operator in terms of inference latency and accuracy within and across the considered tiers?
- RQ4: What is the impact of the Partitioned operator in terms of inference latency and accuracy across the considered tiers?
- RQ5: What is the impact of hybrid operators in terms of inference latency and accuracy within and across the considered tiers?
- RQ6: What is the impact of network bandwidth variations on the deployment strategies in terms of inference latency?

Answering these research questions guides MLOps Engineers and researchers in the AI field to better understand and assess the impact of how and where black-box models are deployed in an Edge AI environment. The results of this paper provide valuable insights for MLOps Engineers debating the most feasible choices for performing inferences in Edge AI contexts.

The assessment in this study simulates an Edge AI deployment architecture for interconnected Mobile, Edge, and Cloud tiers using Docker containers. These containers provide a lightweight and consistent environment for realistic hardware specifications and network conditions. Network conditions, including bandwidths between tiers, are emulated using Linux Traffic Control. AI inference experiments utilize ONNX runtime executors tailored to the hardware limitations of the tiers, processing .onnx models after applying specific operators. Input data consists of larger-sized image samples to evaluate system scalability under computationally demanding scenarios. Accuracy measurements are also performed within each tier using the full validation dataset. This setup enables a comprehensive analysis of inference latency and accuracy across various deployment strategies under realistic conditions.

Our empirical evaluation reveals that black-box deployment strategies significantly impact inference latency and accuracy across Edge AI tiers. We find that hybrid strategies, particularly Quantization + Early Exit on Edge, offer the best latency-accuracy trade-off when a medium accuracy loss is acceptable. When accuracy preservation is paramount, Quantization alone on Edge outperforms other configurations. In resource-constrained mobile environments, Mobile-Edge Partitioning provides preferable latency over full Mobile deployments. Moreover, Cloud deployment becomes effective for small input models even at lower bandwidths ($\leq$ 10 Mbps), whereas larger input models require $\geq$ 50 Mbps for performance parity. Moreover, network bandwidth plays a critical role in shaping optimal deployment strategies.

The rest of this paper is structured as follows. Section 2 discusses the background of the study. Section 3 presents prior works in this field. Section 4 explains the approach, including subjects, experimental

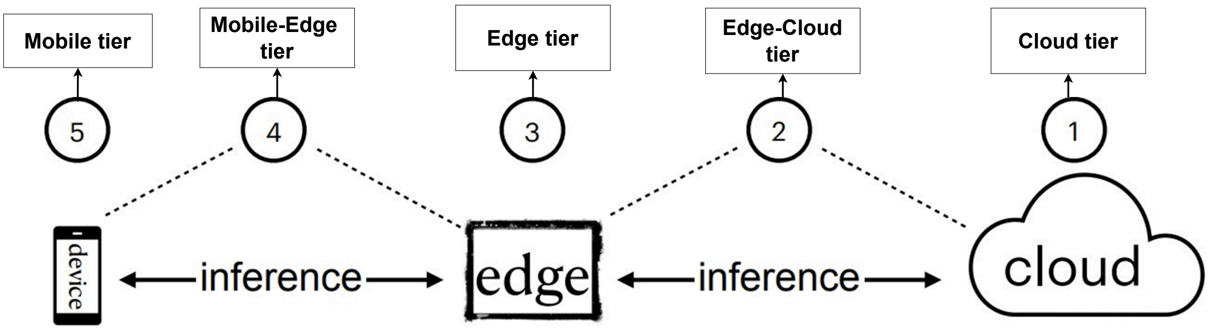


Fig. 1: Graphical overview of Single-tier (Mobile, Edge, Cloud) and Multi-tier (Mobile-Edge, Edge-Cloud) Edge AI deployment Strategies

setup, metrics for evaluating model performance, motivation and approach for each research question, and data analysis. Section 5 describes the results for the five research questions. Section 6 discusses the results and compares the results of the RQs. Section 7 proposes the threats to the validity of the paper, followed by the conclusion in Section 8.

2 Background

2.1 Deep Learning Architecture

The architecture of a Deep Learning (DL) model is composed of layers that transform the input data using mathematical operations to produce an output. Layers are organized sequentially to form a Deep Neural Network (DNN) architecture [55]. The resulting computational graph represents the flow of data and computations through a neural network, i.e., a representation of how the layers are connected and how data moves from one layer to another. Each layer is a node in the graph, and the connections between nodes (Edges) show the data flow. A node encapsulates the entire computation performed by that layer, including all its individual elements. However, the individual elements within a layer, such as neurons, weights, biases, activation functions, and other internal components, are not explicitly represented as separate nodes in the graph. This graph structure allows frameworks to efficiently execute forward and backward passes (propagation) during training and inference processes.

During the training process, the forward pass computes predictions and loss, while the backward pass computes gradients for weight updates to minimize the loss. During the inference process, only the forward pass is used to make predictions or generate output based on input data. Weights are the learnable parameters associated with the layers in the model. These parameters are learned during training to optimize the model's performance. Activations are the intermediate outputs produced by layers as data flows through the computational graph [2].

2.2 Monolithic Edge AI deployment

The traditional deployment of an entire model on Mobile or Cloud (as shown in Figure 1) for inference has some important limitations. For example, deploying the entire AI model on a Mobile device leads to slower computation, whereas on the Cloud it leads to higher transmission latency and potential data privacy threats. Various researchers experiment with ways to deploy models on the Edge (as shown in Figure 1) due to its closer proximity to the Mobile device and faster computational capabilities than a Mobile device [36,64]. However, Edge devices/servers are often smaller and less powerful than centralized servers or Cloud resources, which limits their processing power. To address these limitations, there is a need for alternative deployment strategies (i.e., Model Partitioning, Model Compression, Model Early Exiting) that aim to mitigate these challenges.

2.3 Multi-tier Edge AI Partitioning

The Multi-tier Edge AI deployment strategy partitions (splits) an AI model between Edge-Cloud and Edge-Mobile tiers, respectively, as shown in Figure 1). This overcomes the limitations of traditional deployment strategies by enhancing data privacy, increasing computation efficiency, reducing memory requirements, improving scalability, and providing a flexible architecture that can be easily adapted to different use cases and deployment scenarios.

To illustrate the applicability of this approach, consider a real-time traffic monitoring and incident detection system within a Mobile-Edge-Cloud infrastructure. Cameras mounted on traffic lights, vehicles, and drones continuously capture high-definition video feeds of a busy intersection. These video streams are transmitted to an Edge server located nearby, which processes the footage in real-time using lightweight Computer Vision models. The Edge server detects key objects such as vehicles, pedestrians, and potential road hazards, enabling immediate actions like adjusting traffic signals or issuing warnings for approaching vehicles when pedestrians are detected.

Simultaneously, the Cloud aggregates data from multiple Edge servers across the city, providing a centralized platform for large-scale analysis. By applying Deep Learning (DL) models, the Cloud identifies traffic patterns, predicts congestion zones, and optimizes traffic management strategies. These insights are then used for long-term planning, such as predicting rush hour congestion or adjusting traffic light timings across multiple intersections. This scenario exemplifies how Multi-tier Edge AI Partitioning can optimize resource usage and enhance system performance by distributing computational tasks appropriately across Mobile, Edge, and Cloud tiers.

The key insight is the observation that most of the strategies used to enable Partitioned model inference across, for example, Mobile and Edge devices, operate on the computation graph underlying modern DL architectures [32]. Such graphs express the different elements of neural networks, as well as the calculations performed while traversing edges in the graph. By finding the right edges to split, one can implement model Partitioning, effectively obtaining two or more sub-graphs.

Overall, model Partitioning is a powerful strategy for optimizing the deployment of DNN models on resource-constrained devices. This family of Edge AI operators partitions models across the tiers of an Edge AI Environment:

- Mobile-Edge Partition: The Mobile device receives the input data, performs inference on the first half of the model, and then sends the intermediate output to the Edge device. Afterward, the Edge device performs inference on the second half of the model using the received intermediate output from the Mobile device. Finally, the Edge device sends the final output back to the Mobile device.
- Mobile-Cloud Partition: The Mobile device receives the input data, performs inference on the first half of the model, and sends the intermediate output to the Cloud device via the Edge device. Then, the Cloud device performs inference on the second half of the model using the received intermediate output from the Edge device. Finally, the Cloud device sends the final output back to the Mobile device via the Edge device.
- Edge-Cloud Partition: The Mobile device first receives the input data and sends it to the Edge device as it is. Then, the Edge device performs inference on the first half of the model using the received input data and sends the intermediate output to the Cloud device. The Cloud device further performs inference on the second half of the model using the received intermediate output from the Edge device. Finally, the Cloud device sends the final output back to the Edge device, which in turn sends it back to the Mobile device.

2.4 ONNX Run-time for inference

ONNX (Open Neural Network Exchange)[1] is a standard format built to represent inter-operable AI models that run on a variety of hardware platforms and devices. The core of the ONNX model is the computational graph, which represents the structure and computations of the model. The computational graph consists of nodes that represent individual operations or layers in the model. The nodes take input tensor(s), perform the specified operation, and produce output tensor(s). The nodes have attributes that

[1] https://onnx.ai/

define the type of operation, input and output tensors, and any parameters or weights associated with the operation.

The ONNX Runtime[2] is a high-performance inference Engine for deploying ONNX models to production for real-time AI applications. It is optimized for deployment on various devices (i.e., Mobile, Edge, and Cloud) and hardware platforms (i.e., CPUs, GPUs, and specialized accelerators). We consider ONNX as the subject models' format and the ONNX Run-time as an inference Engine for this study.

ONNX Runtime Execution Providers are a set of plug-ins that enable the execution of ONNX models on a wide range of hardware and software platforms. The ONNX Run-time supported Execution Providers studied and considered as a back-end for hardware acceleration while performing model inference are the CUDA Execution Provider, which uses GPU for computations, and the default CPU Execution Provider, which uses CPU cores for computation.

2.5 Intel Neural Compressor

Intel Neural Compressor (INC) is an open-source toolkit designed to optimize DL models for better performance on Intel hardware. It provides a comprehensive set of features to enable Quantization, Pruning, and other optimizations, making models run faster and consume less power without significant loss in accuracy. INC tool is specifically designed to optimize neural networks for deployment on Intel hardware, such as CPUs or FPGAs.

3 Related Work

In previous surveys [129,110,18,81], eight families of Edge AI model inference operators are discussed: Model Compression (Quantization, Weight Pruning, Knowledge Distillation), Model Partition, Model Early-Exit, Edge Caching, Input Filter, and Model Selection. Moreover, Matsubara et al. [74] conducted a comprehensive survey of the various approaches for Partitioning, Early Exiting, and their combinations with each other and with other operators (such as Bottleneck Injection, Pruning, and/or Knowledge distillation). As summarized in Table 1, many studies focus on in-depth analysis of individual operators or comparing two different operators. Only one study [73] considers three operators, suggesting that there is limited research comparing a larger set of operators within the context of Edge AI. Evaluating three (or more) operators allows for a more comprehensive analysis of their relative performances and trade-offs, which is a focus point of our study.

Among the various kinds of operators, we focus on operators that correspond to model transformations, i.e., modifications in the structure, parameters, or behavior of ML models. This subset of operators can be further categorized into white-box and black-box operators. In previous studies, the white-box operators discussed are Model Pruning and Knowledge Distillation, and the black-box operator discussed is Model Partitioning. Some of the operators like Model Quantization, Model Early Exiting, and Bottleneck Injection can be performed in both black-box and white-box manner. In real-world scenarios, applying white-box operators to ML models requires thorough domain expertise about their internal workings to fine-tune or retrain them, due to which there is a practical need to focus on black-box operators that can optimize models without fine-tuning or retraining. Therefore, we narrowed down our study to the black-box operators (i.e., Model Partition, Model Early Exit, Model Quantization) and their combinations to provide empirical data aimed at understanding how to optimize models robustly and tackle the challenges posed when deploying to heterogeneous Edge AI environments. Among the black-box operators, we excluded the Bottleneck Injection operator as it focuses on intermediate data compression techniques (such as lossless/lossy compression, clipping, etc) and does not inherently perform transformations on the DNN models themselves, which is the goal of our study.

Table 1 compares the distinctive features of our study and previous studies. Among the ten studies focusing on Non-Hybrid Partitioning operators, all of them were performed by black-box transformations, which suggests that this operator requires no retraining or fine-tuning. Among the 25 Early Exit studies, two studies were performed in a black-box manner, which shows that it is feasible to perform Early Exiting without retraining or fine-tuning. Among 21 Quantization studies, the majority of them (16) were performed in a black-box manner, indicating that this approach is more commonly used. In terms

[2] https://github.com/microsoft/onnxruntime

Table 1: Distinctive Features: Prior Work vs. Our Approaches

Ref.	P	E	Q	OO	BT	SS	M	E	C	ILA	ID	TD	RoI	AoA	CaO
Our work	✓V	✓XV	✓XV		✓	✓	✓	✓	✓	S	✓		✓	✓	✓
[120,20,80]	✓				✓	✓	✓	✓		A	✓		✓		
[50]	✓				✓		✓		✓	S	✓	✓			
[43,62]	✓				✓			✓	✓	A	✓		✓		
[48]	✓				✓		✓	✓		S	✓				
[88]	✓				✓		✓	✓	✓	A		✓	✓		
[59]	L		L		✓			✓	✓	S	✓				✓
[127]	✓				✓			✓		A	✓		✓		
[10,25]	✓			BI			✓		✓	S	✓			✓	✓
[24]	✓			BI			✓		✓	S	✓				✓
[14]	✓			BI	✓			✓	✓		✓			✓	✓
[44]	✓			BI				✓		A	✓		✓	✓	✓
[11,4]	✓			BI			✓		✓		✓			✓	✓
[99,47]	✓			BI			✓	✓			✓			✓	✓
[56]	✓			BI				✓	✓	A	✓		✓	✓	✓
[75,76,70,72,121]	✓			BI			✓	✓		S	✓			✓	✓
[97]	✓			BI				✓	✓		✓			✓	✓
[71]	✓			BI			✓	✓		A	✓		✓	✓	✓
[33,34]		✓			✓						✓			✓	
[66]		✓						✓	✓		✓			✓	
[91,111,119,118,108]		✓									✓			✓	
[102,117,23]		✓										✓		✓	
[116]		✓								S		✓	✓	✓	
[128]		✓								S		✓		✓	
[113,9,106]		✓								S	✓		✓	✓	
[85]		✓		PR						S	✓		✓	✓	✓
[82]	I	I				✓		✓		A	✓		✓		✓
[107]	I	I					✓	✓	✓		✓			✓	✓
[123]	I	I					✓	✓		S	✓			✓	✓
[54]	I	I					✓		✓	S	✓		✓	✓	✓
[73]	I	I		BI			✓	✓		S	✓			✓	✓
[90,60]		✓		KD							✓			✓	✓
[65]		✓		KD								✓		✓	✓
[5,7,12,27,29,40,57,77,83,101]			✓		✓						✓			✓	
[30,45,126]			✓		✓						✓	✓		✓	
[61]			✓		✓					NM	✓		✓	✓	
[26,100,96]			✓								✓	✓		✓	
[105]			✓							A	✓		✓	✓	
[39]			✓	PR							✓			✓	✓

[1] P: Partitioning, Q: Quantization, E: Early Exiting, X: Quantized Early Exit, V: Quantized Early Exit Partitioned, L: Combination of Partitioning and Quantization, I: Combination of Partitioning and Early Exiting, OO: Other Operators (BI: Bottleneck Injection, PR: Pruning, KD: Knowledge Distillation), BT: Black-Box Transformations, SS: Simulated Setup (empty cell represents Real or Emulated Hardware Setup), M: Mobile, E: Edge, C: Cloud, ILA: inference Latency Approach (S: sequential, A: Asynchronous, NM: Not Mentioned), ID: Image Data, TD: Textual Data, RoI: Range of Input Sizes (for latency evaluation), AoA: Analysis of accuracy, CaO: Comparison across Operators, Empty cells of Mobile, Edge, and Cloud mean they are not an Edge AI setup

of hardware setup, four of the previous studies consider simulated setup, indicating that it is feasible to consider this kind of setup for testing the operator's performance. In terms of Mobile, Edge, and Cloud tiers, only two studies consider all three tiers, suggesting that this area of research is relatively less explored. In terms of Mobile, Edge, and Cloud tiers, ILA, and RoI, no study considers sequential inference of a range of inputs (with varying sizes) in an Edge AI setup (Mobile, Edge, and Cloud tiers), which was explored in our study. In terms of input data, the majority of studies (64) consider image data, indicating that this type of data is more commonly used for the mentioned operators.

In Table 1, there are limited studies (only two) that consider all three tiers (i.e., Mobile, Edge, and Cloud) for Edge AI setup in their experiments. Considering all three tiers collectively provides a more holistic view of real-world deployment scenarios with varying computational and network conditions. Therefore, we considered all three tiers to ensure a comprehensive and versatile approach in our Edge AI setup. Four prior studies [120,82,20,80] employ a simulated Edge AI setup instead of real hardware, the former is more cost-effective and more accessible than real hardware, while also providing a controlled environment, making it easier to isolate and micro-benchmark the latency performance of individual operators. On the other hand, while simulations can closely approximate the behavior of real hardware, they may not replicate all the nuances and complexities of a real-world environment such as hardware/network variability, power consumption, and real-time constraints.

Among the studies considering interconnected multi-tier networks, the majority (20) consider sequential inference in comparison to asynchronous inference (i.e., 9) for latency evaluation across the tiers. In sequential inference, the inference tasks proceed in a step-by-step manner across the tiers of the network and are dependent on each other (i.e., the next inference task waits for the completion of the previous inference task). In asynchronous inference, the inference tasks across the tiers are performed concurrently or independently from each other. We considered sequential inference as it allows us to isolate the performance characteristics of individual operators in a controlled environment. In other words, it allows for uncovering a more deterministic impact of input data sizes and network/computational resources of heterogeneous Mobile, Edge, and Cloud tiers on the operators's latency performance, similar to how micro-benchmarks operate.

The majority of the prior studies focus on image data instead of textual or speech data as input for inference of these black-box operators due to its prevalence in real-time deployment scenarios. Image data is often more complex and less interpretable than text or speech data, requiring more bandwidth for transmission and more storage space than text or speech data typically due to its larger size. It requires more computational resources to process possibly due to their usage in computationally intensive tasks like object detection, image classification, and image segmentation. As a result, this is why CV models are commonly studied in prior work, requiring image data as input for inference. As such, we narrowed down the scope of the data and model in our study to the CV domain.

To conclude, our paper provides a novel empirical study of Edge AI deployment strategies, which are mappings of the black-box operators (i.e., Partitioning, Early Exiting, Quantization), and their combinations, onto Edge AI tiers (i.e., Mobile, Edge, and Cloud), and their combinations, to analyze the optimal trade-off in terms of latency and accuracy in real-time deployment scenarios. The previous studies, as mentioned in Table 1, combine Partitioning with either the Early Exiting [82,107,123,54,73] or Quantization operator [59]. In our study, we went one step further and analyzed unexplored combinations among these three operators, like Quantized Early Exit and Quantized Early Exit Partitioned. To our knowledge, there is no comprehensive study in previous work (Table 1) on the comparative analysis of these three black-box operators and their specific combinations in the context of Edge AI to decide which operator is optimal in which deployment scenario. Secondly, our study in comparison to previous studies (Table 1), evaluates the end-to-end (round trip) latency of the deployment strategies in an Edge AI setup (Mobile, Edge, and Cloud tiers). The third contribution is our focus on measuring the latency of deployment strategies across a wide range of varying input (i.e., image) sizes using sequential inference requests (which have not been explored in previous studies). This contribution helps in analyzing the impact of input data on the proposed deployment strategies.

Below, we discuss existing work related to the three operator families which we considered in our study.

3.1 Partitioning

As explained in Section 2.3, the Model Partitioning operator performs black-box transformations splitting a given model into head (1st half Partition) and tail (2nd half Partition) sub-models at a Partition point such that the two sub-models, when feeding the output of the head into the input of the tail, produces the same output as the original model. While in some studies [10,14], the Partitioning point is chosen heuristically, in the majority of studies performing Model Partitioning [24,48,50,59,123,88,20,120,80,62, 43], various factors like computational load, network cost, energy consumption, input data sizes and/or privacy risk are evaluated for each of the Partitioning points of the DNN models during deployment across the Edge AI environment to inform the selection. There is no generalized optimal Partition point, as it varies for models with different architectures [74]. Therefore, in our study, we simplified our approach by considering equal-size (MB) Partitioning to do a fair evaluation of each of the subjects considered in our study.

Many of the CV models (i.e., AlexNet, VGG 11/16/19, DenseNet, ViT, NiN, ResNet 18/34/50/56, GoogLeNet, AgeNet, GenderNet, Inception-v3, BNNs, eBNNs) considered in previous studies [120,20, 80,50,62,43,48,88,59,127,10,24,107,123,54] for model Partitioning based on black-box transformations have weak accuracy performance and model complexity within reach of resource-constrained Mobile tiers. However, in our study, more accurate and complex state-of-the-art CV models are considered (such as Wide ResNet-101, ResNext-101, FCN, and DUC) to analyze the latency vs accuracy trade-off in an Edge AI environment with heterogeneous Mobile, Edge, and Cloud tiers.

The Bottleneck Injection (BI) operators have also been previously studied in combination with Model Partitioning to reduce the transmission, computation, and energy costs across the Mobile, Edge, and Cloud tiers. These introduce artificial bottlenecks to DNN models by compressing the intermediate data, modifying the DNN architecture, and/or both. The Bottleneck Injection techniques that do not involve re-training of DNN models include Intermediate Data Compression using Quantization, Tiling, Clipping, Binarization, Entropy Coding, and Lossy/Lossless Compression, which are analyzed in a few previous studies [10,24,14]. However, most of the BI operators require extensive re-training of models as they modify the DNN architectures with Auto Encoders [25,44,11,99,56,47,121,4,97,71,75,76], Head Network Pruning [47], and Head Network Distillation [70,72,73,4,97,71,75,76]. The mentioned black-box and white-box BI operators may affect the accuracy performance due to intermediate data compression and architectural modifications, respectively.

In our study, we treat the Partitioning operator as a black-box transformation of a model into two sub-models (this number is commonly used in previous studies) that do not modify the input or intermediate representations in the models (i.e., the final output will not change), hence, preserving the accuracy. This is important, because the Bottleneck Injection operators can be costly and time-consuming (especially the ones involving architectural modifications), and would change the known, possibly certified behavior of an existing model.

3.2 Partitioning Approach

We considered the simple, but effective aspect of equal-size (MB) Partitioning to perform a fair evaluation of the subjects. Equal-sized Partitioning allows each Partitioned model to have a similar level of complexity and workload, which can help to balance the computational load across the tiers of the Edge AI Environment, ignoring the heterogeneity of tiers. This fairness can be crucial for assessing the models objectively and avoiding biases introduced by variations in computational capabilities across tiers or certain tiers being underutilized or overloaded, promoting efficient resource utilization. Moreover, due to the variation in the structural properties of the subjects considered for CV tasks, this straightforward approach might lead to a larger amount of intermediate data being transferred to the Mobile, Edge, and Cloud network.

For each of the subjects (i.e., ResNet, ResNext, FCN, DUC) considered in our study, we observed that the size of the sub-graphs of the computational graph gradually increases while traversing from the input to the output node. As we progress through the sub-graphs, the receptive field (sensitive to the region of the input image) of nodes expands, incorporating information from a larger context, which often increases the size of the sub-graphs. This observation is consistent with the typical architecture of DNNs, where lower layers capture low-level features, and higher layers combine these features to form more complex representations. Therefore, choosing the Partition point(s) closer to the end might balance the size between the two sub-models, as shown in the Partitioning examples in the Appendix.

So, with that in mind, Algorithm 1 inspects the ONNX computational graph of subjects traversing from the end and heuristically selects the Partition point, i.e., node connection(s), that splits the model into two nearly equal-sized sub-models. This algorithm is different from the Early Exit algorithm 2 because Partitioning focuses on finding the node connection(s) that can Partition the model into equal sizes, which requires manually checking the sizes of Partitioned models in our study (proof-of-concept algorithm). It deviates from the Early Exit approach, in which we find and skip identically structured sub-graphs by analyzing their structures and input/output node dimension in ONNX computational graphs.

The models were partitioned so that the connection(s) used for doing that connect the outputs of the 1st half-Partitioned model to the input of the 2nd Half-Partitioned model. For ResNet and ResNext, the connection as shown in Figure 18, available at the start of the 9th sub-graph (from the end), was used as the Partitioning point, as it showed the lowest difference in size between the sub-models (i.e, 7-8 MB) (based on line 5 to 16 in algorithm 1). These sub-graphs are structured in a way such that one of its branches contains the interconnected nodes and the other branch is a connection (without any interconnected nodes) that eventually merges at the end of the sub-graph. It is not feasible to use the connections within the branches of these sub-graphs for Partitioning as it requires each of these branches to have inter-connected nodes to cut the sub-graph.

The architecture of FCN consists of two main branches that extend from the input node and merge at the last sub-graph located at the end. One of the branches consists of all the heavy-weight sub-graphs and the other branch consists mainly of an elongated connection, i.e., a connection that spans across the network without containing complex computations or heavy-weight sub-graphs (line 22 to line 23 in Algorithm 1). These two side branches merge into a sub-graph at the end, which computes the final output. Therefore, the Partitioning of FCN requires cutting the connection within these two branches to maintain the information flow, as shown in Figure 19. The first connection is situated at the start of the forth sub-graph from the end within the heavy-weight side branch, and the other connection (within the lightweight side branch) is situated just before the merge (line 24 in Algorithm 1). The variation in FCN's Partitioned model sizes is limited to 3MB (based on line 5 to 16 in Algorithm 1). As explained earlier (for ResNet/ResNext), these sub-graphs lack interconnected nodes in one of their branches, making it unfeasible to use their branches' connections for Partitioning.

For DUC, the connection situated at the start of the forth and third sub-graph from the end shows sub-models with 37 MB and 9MB variation, respectively (line 29 in Algorithm 1). The forth sub-graph from the end contains inter-connected nodes within both its branches and, therefore, cutting the connection within each of its branches can lead to an even more effective balance in size for DUC's sub-models. Therefore, we explored the connections within the branches of this sub-graph and Partitioned the DUC model at the connections shown in Figure 20, which resulted in sub-models with around 1MB size variation (based on line 30 to 40 in Algorithm 1). This shows that the challenges associated with equal-size Partitioning might vary for models with different architectures and therefore manual analysis of their computational graphs and Partitioned sub-model sizes is beneficial in such cases.

The manual splitting of the subject's Identity models into two equally sized (in MB) sub-models is performed using ONNX Python APIs. We use the extract_model() function on each of the Identity models to perform the following tasks: 1) Extract 1st half of the Partitioned model by traversing from the input node to the Partition point; 2) Extract 2nd half of the Partitioned model by traversing from the Partition point to the output node. At the Partition point, the output tensor name(s) of the 1st half-Partitioned model is identical to the input tensor name(s) of the 2nd half-Partitioned model. The 1st half-Partitioned sub-model is used for inference in either the Mobile or Edge tier, while the 2nd half-Partitioned sub-model is used for inference in either the Edge or Cloud tier.

3.3 Early Exiting

Early Exiting is another family of Edge AI deployment operators allowing DNN models to make early predictions without having to wait until the entire computation process (full forward pass) is completed by terminating the execution at Early Exits/Classifiers/Sub-branches [98]. The benefits of model Early Exiting include faster inference speed, reduced energy consumption, and increased model efficiency at the cost of lower accuracy performance. Implementing Early Exiting on a trained model improves the model's runtime performance, especially if the model was initially designed for high accuracy and not optimized for efficiency.

In the majority of the previous studies, the Early Exits require re-training of the base models either by joint training or separate training. In Joint training, the (Early) Exits are trained simultaneously with a model [23,54,66,91,102,106,107,108,111,116,118,119,123,128] by defining a loss function for each of the classifiers and minimizing the weighted sum of cross-entropy losses per sample. In contrast, in separate training [65,73,117,113], model training is performed in the first stage then the training of the Early Exits is performed, such that the pre-trained model parameters are frozen. There are some studies [33, 34] that perform Early Exiting by comparing the output of an Early Exit with the corresponding class means using Euclidean distance. If the output of an Early Exit is not close enough to a class mean, the execution continues and the same process is performed for the next Early Exit in the DNN model. In other words, the Early Exiting is performed dynamically at inference in a black-box manner.

In our study, we achieve a similar effect of Early Exiting by performing manual and static modifications on the computational graphs of the subject models to short-circuit (skip) the similarly structured

Algorithm 1 Equal-Size Partitioning of Subject models

Input: ONNX computational graph G_{subject} for each subject model
Output: Equal-size Partitioned sub-models
Initialize Min Size Difference $\Delta_{\min} \leftarrow \infty$
Initialize Partition Point $P \leftarrow$ null
function PARTITION(G_{graph}, $\Delta_{\min}$, P)
  Reverse traverse G_{graph} starting from the end
  while not reached the beginning of G_{graph} **do**
    Use connection at the start of the current sub-graph S_{current} for Partitioning G_{graph}
    Calculate size difference Δ of Partitioned sub-models
    if $\Delta < \Delta_{\min}$ **then**
      $\Delta_{\min} \leftarrow \Delta$
      $P \leftarrow$ connection at the start of S_{current}
    end if
  end while
  return $P, \Delta_{\min}$
end function
function PARTITIONRESNE(X)T(G_{subject}, $\Delta_{\min}$, P)
  $P, \Delta_{\min} \leftarrow$ Partition(G_{subject}, $\Delta_{\min}$, P)
  return Partitioned sub-models of G_{subject} using P based on $\Delta_{\min}$
end function
function PARTITIONFCN(G_{subject}, $\Delta_{\min}$, P)
  Select heavy-weight side branch $G_{\text{heavy-weight}}$
  Select light-weight side branch $G_{\text{light-weight}}$
  $P \leftarrow$ Use one connection within the $G_{\text{light-weight}}$ just before the merge for Partitioning G_{subject}
  $P, \Delta_{\min} \leftarrow$ Partition($G_{\text{heavy-weight}}$, $\Delta_{\min}$, P)
  return Partitioned sub-models of G_{subject} using P based on $\Delta_{\min}$
end function
function PARTITIONDUC(G_{subject}, $\Delta_{\min}$, P)
  $P, \Delta_{\min} \leftarrow$ Partition(G_{subject}, $\Delta_{\min}$, P)
  Reverse traverse G_{subject} starting from the end
  while not reached the beginning of G_{subject} **do**
    if the side branches of the current S_{current} contain inter-connected nodes **then**
      Select connections within the branches of the S_{current} for Partitioning
      Calculate size difference Δ of Partitioned sub-models
      if $\Delta < \Delta_{\min}$ **then**
        $\Delta_{\min} \leftarrow \Delta$
        $P \leftarrow$ connections within the S_{current}
      end if
    end if
  end while
  return Partitioned sub-models of G_{subject} using P based on $\Delta_{\min}$
end function

graph computations. The motivation behind this approach lies in the flexibility and customization it offers in the ONNX framework to MLOps Engineers.

3.4 Early Exit Approach

In our study, the Early Exit process involves modifying the architecture of the pre-trained subjects to include intermediate outputs and adding the necessary logic to allow the models to Exit Early at an intermediate stage in the neural network, where this stage includes intermediate outputs that can be used for prediction. We create Early Exit models by manually terminating the model early using ONNX python APIs [3]. Since the Early Exit mechanism damages the accuracy of inference, a relatively slower Early Exit near the end of the DNN will gain better accuracy performance [106,123]. For this reason, we traverse the ONNX computational graphs for all subjects in reverse order (line 4 to 6 in Algorithm 2), i.e., starting from the end (output node) to check the sub-graphs having identical structures, then short-circuiting them to create an Early Exit. Here, the sub-graph denotes a branch network of graphical nodes.

This process takes into account specific considerations related to the model architecture of subjects and the desired trade-off between accuracy and inference speed. When dealing with DNNs, each sub-graph

[3] https://github.com/onnx/onnx/blob/main/docs/PythonAPIOverview.md

may have specific requirements for the dimensionality of its input and output nodes. If the subsequent sub-graphs have different structures and dimensions, skipping them could lead to incompatible input/output configurations, disrupting the overall flow of the computational process. In our case, the skipping of sub-graphs on each of the subject architectures is based on their identical structure and input/output node dimension. For each of the subjects, we skipped two identical consecutive sub-graphs while traversing from the end (line 10 to 13 in Algorithm 2). Skipping more than two was not feasible due to the variation in structure as well as the input/output node dimension of the sub-graphs preceding them (line 14 to 15 in Algorithm 2). Skipping only one sub-graph might not yield a significant speedup, as the reduction in model size would be limited to 1.07x to 1.12x of the Identity models (line 15 to 17). However, by skipping two consecutive sub-graphs, the model size can be reduced by 1.15x to 1.27x relative to the Identity models (line 15 to 17), resulting in a more substantial speedup during inference. In the architectures of these subjects (refer to example Figure 15 in appendix), there are other sub-graphs (more than two) having identical structures. Skipping them could result in higher latency as they are placed at the early stages of the graph, which are lighter in weight compared to the ones we selected. For them, skipping a higher number of sub-graphs would be required to yield significant model size reduction and faster latency performance. This would result in significant accuracy loss as these sub-graphs are positioned at the earlier stages of the graph, which capture fundamental features or representations of the input data, which are essential for accurate predictions.

Concretely, on the Identity model of each subject, Algorithm 2 first uses the extract_model() function to perform the following tasks: 1) Extract an Early Exit sub-graph having an input node at the start and an Early Exit point at the end; 2) Extract a decision sub-graph having the Early Exit point at the start and the output node at the end. The purpose of the decision sub-graph is to make the final prediction based on the information available up to the Early Exit point. Then, we use the merge_models() function to merge the Early Exit sub-graph with the decision sub-graph. This effectively allows us to either execute the entire model (i.e., the Identity model) or to skip the last two consecutive and identical sub-graphs (in terms of structure and input/output node dimension), essentially Exit the model halfway. We provided graphical illustrations of the Early Exit operation on the ONNX computational graphs of the subjects in Figure 15 16 17.

Algorithm 2 Early Exit based on Sub-Graph Similarity

1: **Input:** ONNX computational graph for each subject model
2: **Output:** Modified computational graph with Early Exit
3: Initialize Max Size Difference $\Delta_{\max} \leftarrow 0$
4: **for** each subject model **do**
5: Extract the ONNX computational graph G_{subject}
6: Reverse traverse G_{subject} starting from the end
7: Initialize skip count $count \leftarrow 0$
8: **while** not reached the beginning of G_{subject} **do**
9: Extract current sub-graph S_{current}
10: Extract preceding sub-graph $S_{\text{preceding}}$
11: **if** S_{current} has identical structure and input/output dimensions as $S_{\text{preceding}}$ **then**
12: Remove S_{current} from G_{subject}
13: Increment *count* by 1
14: Store the updated graph as $G_{\text{early\_exit}}$
15: Compute size difference $\Delta \leftarrow \text{Size}(G_{\text{subject}}) - \text{Size}(G_{\text{early\_exit}})$
16: **if** $\Delta > \Delta_{\max}$ **then**
17: Update $\Delta_{\max} \leftarrow \Delta$
18: **end if**
19: **else if** S_{current} does not have identical structure and input/output dimensions as S_{previous} **then**
20: **Break**
21: **end if**
22: **end while**
23: **end for**

3.5 Quantization

Quantization [35] is a member of the Model Compression family of Edge AI deployment operators, where the neural network's calculation reduces from full precision (i.e., 32-bit floating point format) to reduced precision (e.g., 16-bit, 8-bit integer point format) to decrease both the computational cost and memory footprint, making inference more scalable on resource-restricted devices [51]. As explained in previous literature surveys [31, 114], there are two popular Quantization approaches used in machine learning to optimize DNNs for deployment on hardware with limited numerical precision, such as CPUs, GPUs, and custom accelerators, i.e., QAT, and PTQ.

QAT incorporates Quantization into the training process itself. This is done using techniques such as fake Quantization or simulated Quantization, which simulate the effects of Quantization on the weights and activations during the forward and backward passes of training. The QAT Quantization method involves training data and back-propagation for its fine-tuning process, which requires a full training pipeline, which takes significant extra training time and can be computationally intensive when dealing with large and complex neural networks [26, 100, 96, 105, 39]. In particular, the standard forward/backward passes are executed on a model that uses floating-point precision, and the model parameters are Quantized after each gradient update. By training the model to be more robust to Quantization, QAT results in models that are more accurate after Quantization than PTQ. QAT typically involves two stages: calibration, where the appropriate range of values for the weights and activations is determined, and fine-tuning, where the model is trained with the Quantized weights and activations.

An alternative to the more resource-intensive QAT method is PTQ. PTQ involves reducing the weights and activations of a pre-trained model to lower integer bits, all without the need for fine-tuning (i.e., in a black-box manner) [5, 7, 12, 27, 29, 40, 57, 77, 83, 101, 30, 45, 126, 61]. This can be done using techniques such as static or dynamic PTQ. In dynamic PTQ, the Quantization parameters are dynamically calculated for the weights and activations of a model during runtime and are specific for each inference, while for static Quantization, the Quantization parameters are pre-calculated using a calibration data set and remain static during each inference. The advantage of PTQ lies in its low and often negligible overhead. Unlike QAT, which relies on a substantial amount of labeled training data for retraining, PTQ is advantageous in scenarios where data is limited or unlabeled. Moreover, we observed that 16 studies considered the black-box Quantization (PTQ) and only 5 studies focused on white-box Quantization (QAT) as shown in Table 1. This suggests that black-box Quantization is more common and is therefore considered for evaluation in our study. We keep the pipeline more straightforward by performing Static PTQ Quantization, which just requires representative data (i.e., validation set in our study) to compute statistics such as mean and standard deviation of weights and activations.

3.6 Quantization Approach

In our study, we used an INC with ONNX Runtime (CPU) backend to perform static PTQ on the subjects. Static PTQ uses a calibration dataset to determine the Quantization parameters, such as scaling factors and zero points for the model. These parameters are essential for representing the floating-point weights and activations of a model in lower-precision fixed-point formats, which are required for Quantization. The calibration dataset is used to represent a representative subset of the input data that the model is likely to encounter during inference. For each subject, its validation set is passed as the calibration data to capture the data distribution and help identify appropriate Quantization parameters for the model to maintain the desired level of accuracy.

The main advantage of using this technique is that it can lead to a significant reduction in memory requirements and computation time while still maintaining model accuracy. This is especially important in scenarios where the model needs to be deployed on resource-constrained devices, such as Mobile or Edge devices. In static PTQ, the weights and activations of a pre-trained model are Quantized to a fixed precision (i.e., 8-bit integers) by the INC.

3.7 Hybrid Approach

The RQ5 evaluates the impact of combining the three deployment operators (i.e., Quantization, Early Exit, and Partitioning). We perform Early Exit on the Quantized models by skipping identically structured sub-graphs from the end of the ONNX computational graphs, which is identical to the approach of Early Exit in Identity models, as explained in RQ3 (Section 3.4). We provide graphical illustrations of the Quantized Early Exit operation on the ONNX computational graphs of the subjects as shown in Figures 21 22 23.

We manually Partition the Quantized Early Exit models into two nearly equal-sized sub-models to generate the Quantized Early Exit Partitioned operator using a similar procedure as for the Partitioning of Identity models in RQ4 (Section 3.2). Here, the second half-Partitioned model contains the Early Exit operation, allowing it to make early predictions. The first-half and second-half Partitioned sub-models are used for inference in Mobile/Edge and Edge/Cloud tier, respectively. We provided graphical illustrations of the Quantized Early Exit Partitioned operator on the ONNX computational graphs of the subjects as shown in Figures 24, 25, and 26.

4 Methodology

This section presents the methodology adopted to address the research questions (RQs) introduced earlier. Our approach is grounded in the Goal/Question/Metric (GQM) paradigm [6], which provides a structured framework for defining measurement goals, formulating relevant research questions, and identifying appropriate metrics to assess the outcomes. By aligning our methodology with the GQM model, we ensure that our evaluation is systematic, goal-driven, and traceable from high-level objectives to concrete measurements.

4.1 Goal, Research Questions, and Metrics

The goal of the experiment is to analyze Edge AI deployment strategies to evaluate their impact on latency and accuracy performance from the perspective of MLOps engineers. In this context, a deployment strategy refers to a combination of operators and tiers. The operators include Identity, Quantization, Early Exit, Quantized Early Exit, and Quantized Early Exit Partition. The tiers include Mobile, Edge, Cloud, Mobile-Edge, Edge-Cloud, and Mobile-Cloud.

Table 2 summarizes how our overarching goal maps to each research question and the associated evaluation metrics.

4.1.1 RQ1: What is the impact of Monolithic deployment in terms of inference latency and accuracy across the considered tiers?

Motivation This question aims to empirically assess the possible differences in terms of inference performance between the three tiers (i.e., Mobile, Edge, and Cloud) during the Monolithic deployment of Identity models (i.e., models to which the Identity operator is applied, meaning no modification or optimization is performed on the original model). In our study, Monolithic deployment on each tier involves deploying an entire model, along with any necessary pre-processing, post-processing, and inference logic, as a single unit. The goal of this research question is to analyze the impact of factors like computational resources, network bandwidth, and input data on inference latency for the three Monolithic deployment scenarios. The inference accuracy of Identity models is also computed as the baseline for analyzing the performance in later RQs.

4.1.2 RQ2: What is the impact of the Quantization operator in terms of inference latency and accuracy within and across the considered tiers?

Motivation This question evaluates the impact of the Quantization operator through two key comparisons. First, we compare the inference latency and accuracy of Quantized models against Identity models within the same deployment tier (i.e., Mobile, Edge, and Cloud). Second, we analyze the effect of Quantization across the three Monolithic tiers to examine its behavior in different deployment environments. These comparisons are designed to empirically explore the trade-offs introduced by Quantization, particularly its potential to reduce latency while maintaining acceptable accuracy, thereby informing its suitability for deployment in resource-constrained versus more capable environments.

Table 2: GQM Mapping of Goals, Research Questions, and Metrics

Goal	Research Question	Metrics
Assess the baseline performance of Identity models deployed monolithically across Mobile, Edge, and Cloud tiers.	**RQ1**: What is the impact of Monolithic deployment (Identity models) in terms of latency and accuracy across tiers?	Inference latency (ms), Top-1 accuracy (baseline)
Evaluate how Quantization affects inference latency and accuracy within and across deployment tiers.	**RQ2**: What is the impact of Quantization on latency and accuracy within and across tiers?	Inference latency (ms), Top-1 accuracy, relative latency reduction (%)
Analyze the effect of Early Exit on latency and accuracy in different deployment environments.	**RQ3**: What is the impact of Early Exit on latency and accuracy within and across tiers?	Inference latency, Top-1 accuracy, number of early exits taken
Examine the effectiveness of Partitioning in reducing latency across multi-tier setups.	**RQ4**: What is the impact of Partitioning across tiers?	Inference latency, Latency difference vs. Monolithic deployments
Assess the impact of hybrid operators combining Quantization, Early Exit, and Partitioning on latency and accuracy.	**RQ5**: What is the impact of hybrid operators (Quantized Early Exit, Quantized Early Exit Partitioning)?	Inference latency, Accuracy, Trade-off analysis (latency vs. accuracy)
Determine how network bandwidth variation influences the latency of deployment strategies.	**RQ6**: What is the impact of bandwidth variation on deployment strategies?	Inference latency under 1, 10, 50, 100, 150, and 200 Mbps

4.1.3 RQ3: What is the impact of the Early Exit operator in terms of inference latency and accuracy within and across the considered tiers?

Motivation To evaluate the impact of the Early Exit operator, we conduct two types of comparisons. First, we compare Early Exit models with their Identity counterparts within the same deployment tier (i.e., Mobile, Edge, and Cloud) to assess how introducing early exits affects latency and accuracy under similar resource constraints. Second, we analyze how the Early Exit operator performs across the three Monolithic tiers to understand its behavior in varying deployment environments. These comparisons are motivated by the need to understand whether Early Exit can effectively reduce inference latency while maintaining acceptable accuracy across different system configurations.

4.1.4 RQ4: What is the impact of the Partitioning operator in terms of inference latency and accuracy across the considered tiers?

Motivation This research question compares the inference latency of multi-tier Partitioning strategies (i.e., Identity models partitioned across Mobile-Edge, Edge-Cloud, and Mobile-Cloud) with that of Monolithic Identity deployments (i.e., complete deployment of Identity models on Mobile, Edge, or Cloud). Additionally, the Partitioned operator in the Mobile-Edge tier is compared with other operators—Identity, Quantization, and Early Exit—in the Edge tier. This tier selection is based on their superior latency performance relative to other combinations (e.g., Edge-Cloud or Mobile-Cloud for Partitioning, and Mobile or Cloud for other operators).

These comparisons aim to empirically assess the effectiveness of the Partitioning operator in reducing inference latency within Edge AI environments and to understand its relative benefits over monolithic and alternative deployment strategies.

4.1.5 RQ5: What is the impact of hybrid Operators in terms of inference latency and accuracy within and across the considered tiers?

Motivation This question compares the inference latency and accuracy performance of combined optimization strategies involving Quantization, Early Exit, and Partitioning. First, the Quantized Early Exit (QE) hybrid operator is evaluated against Non-Partitioned operators (i.e., Identity, Quantization, and Early Exit) within each Monolithic tier (Mobile, Edge, and Cloud). Second, the Quantized Early Exit Partitioned (QEP) operator is analyzed across Multi-tier setups (i.e., Mobile-Edge, Edge-Cloud, and Mobile-Cloud) and compared with the QE operator deployed monolithically. Additionally, the QEP operator in the Mobile-Edge tier is compared with Non-Hybrid operators (i.e., Identity, Quantization, and Early Exit) in the Edge tier. These tier choices reflect configurations that show the most promising latency benefits based on prior observations (e.g., Mobile-Edge for QEP and Edge for non-hybrid operators).

To assess performance degradation, the accuracy of QE is also compared against Non-Hybrid operators. Identity and QE were selected for Partitioning in this and the previous RQ (RQ4) due to their representing the extremes of model size, enabling clearer latency contrasts across tiers. The decision also aligns with prior studies that explored combining Partitioning with Quantization and Early Exit, as discussed in Section 3.

These comparisons aim to empirically evaluate how integrating multiple model optimization operators affects inference latency and accuracy in diverse deployment environments.

4.1.6 RQ6: What is the impact of bandwidth variations on the deployment strategies in terms of inference latency?

Motivation This question aims to empirically assess the impact of network bandwidth variations on the inference latency performance of various deployment strategies, including Identity, Quantization, Early Exit, Partitioning, and hybrid operators, across the Mobile, Edge, and Cloud tiers.

To study this, we focus on experiments involving a single input data sample, where bandwidth is treated as a key independent variable. We evaluate six commonly observed bandwidth levels—1 Mbps, 10 Mbps, 50 Mbps, 100 Mbps, 150 Mbps, and 200 Mbps—across both Mobile-Edge and Edge-Cloud connections. These values were selected to reflect a range of realistic network conditions, from constrained (1 Mbps) to ideal (200 Mbps) scenarios, as typically encountered in mobile and edge computing environments.

This setup allows us to isolate and measure the direct effect of bandwidth on inference latency under each deployment strategy. By varying only the bandwidth while holding other factors constant, we can better understand how network limitations influence latency and which deployment strategies are more resilient to such variations.

4.2 Subjects Selection

A set of four pre-trained, state-of-the-art models from the ONNX Model Zoo [4] and Pytorch Imagenet models store [5] is used as a suitable and representative sample of subjects for the experiment. Testing too many models can be computationally expensive and time-consuming, especially when involving techniques like Quantization, Early Exit, Partitioning, and their combinations. Previous benchmarking studies [38] use a limited set of widely recognized models to evaluate methods. Therefore, using four models aligns with this convention, providing sufficient statistical insights without overburdening the study and strike a balance between variety and manageability. In this study, we focus on computer-vision tasks as they often demand significant computational and network bandwidth resources during deployment in real-world scenarios. To support the generality of the results, we ensure that the models are heterogeneous in terms of architecture, size, scope, and data set. Furthermore, to obtain findings on realistic models, we selected four large and complex image classification and segmentation models, as shown in Table 3.

Image classification is a type of machine learning task where the goal is to assign a label or category to an input image. This is achieved by training a model on a dataset of labeled images, then using that model to predict the labels of new, unseen images [68]. Image segmentation is the process of dividing an image into multiple segments or regions. The goal of image segmentation is to simplify and/or change the representation of an image into something more meaningful and easier to analyze. The output of image segmentation is a set of segments that collectively cover the entire image or a set of contours extracted from the image [79].

[4] https://github.com/onnx/models
[5] https://pytorch.org/vision/main/models.html

The ILSVRC (ImageNet Large Scale Visual Recognition Challenge) dataset is used for evaluating the performance metrics of both the ResNet and ResNext subjects as it is widely used for training and evaluating image classification models. As the network architecture of the ResNet and ResNext subjects is similar, we use different versions of pre-trained weights for ResNet (i.e., IMAGENET1K_V2) and ResNext (i.e., IMAGENET1K_V1) from the torchvision package to obtain a better generalization of our results. We use the COCO (Common Objects in Context) and Cityscapes datasets for evaluating the performance metrics of the FCN and DUC subjects, respectively. For the Image Classification subjects, we exported the ResNet and ResNext models from torchvision.models subpackage to ONNX using the torch.onnx.export() function. For the Image Segmentation subjects, we use the models from ONNX Model Zoo.

Table 3: Subjects of the experiment

Model Name	**Model Size**	**Parameters**	**Scope**	**Dataset**
ResNet [122]	484MB	126.81M	Image Classification	ILSVRC 2012 [95]
ResNext [115]	319MB	83.35M	Image Classification	ILSVRC 2012 [95]
FCN [67]	199MB	51.89M	Image Segmentation	COCO 2017 [63]
DUC [109]	249MB	65.14M	Image Segmentation	CityScapes [16]

4.3 Experimental Variables

The experiment is structured using a factorial design comprising multiple independent and controlled factors. Following guidelines from Wohlin et al. [112], we distinguish between *independent factors*, which are actively varied during the experiment, and *controlled factors*, which are held constant to isolate the effects of the independent ones. The responses (dependent variables) are latency and accuracy.

Independent Factors The primary independent factors include:

- **Operator Configuration (5 levels)**: Quantization, Early Exit, Partitioning, Quantized Early Exit, and Quantized Early Exit Partitioned.
- **Deployment Tier (6 levels)**: Mobile, Edge, Cloud, Mobile-Edge, Edge-Cloud, and Mobile-Cloud.
- **Network Bandwidth (6 levels)**: 1, 10, 50, 100, 150, and 200 Mbps (varied in experiments with single input samples; fixed in others as described below).

In experiments involving multiple input samples, bandwidth is treated as a controlled factor and fixed at 200 Mbps for Mobile-Edge and 1 Mbps for Edge-Cloud tiers to avoid confounding effects.

Controlled Factors (Fixed During the Experiment) Controlled factors are variables that are deliberately kept constant across all experimental conditions to maintain internal validity and ensure that variations in the dependent variables are attributable only to the independent factors. These include:

- **Model Architecture**: The same subject models (e.g., ResNet/ResNeXt, FCN, DUC) are used across all configurations.
- **Input Data and Preprocessing**: Identical datasets and input transformations are applied across all experimental runs.
- **Deployment Tools**: The ONNX runtime and Intel Neural Compressor are consistently used across all experiments.
- **Hardware Configuration**: Hardware specifications (CPU, GPU, RAM) for Mobile, Edge, and Cloud tiers are fixed. CPU for Mobile/Edge: Intel(R) Xeon(R) E7-4870 2.40GHz, CPU for Cloud: Intel(R) Xeon(R) Platinum 8268, GPU: NVIDIA A100 GPU, RAM for Mobile: 4GB, RAM for Edge: 16GB, RAM for Cloud: 64GB RAM.

These controlled factors are considered constant *design parameters* in the experiment and not varied across treatment conditions.

Dependent Variables (Responses) The measured outcomes are:

- **Inference Latency**: Time in milliseconds to complete an end-to-end inference request, including preprocessing, model computation, post-processing, and data transmission.
- **Inference Accuracy**: Evaluated using Top-1/Top-5 accuracy for classification models, and mIoU for segmentation models.

4.3.1 Inference Latency

We define the inference latency as being based on the sum of pre-processing latency, model computational latency, post-processing latency, and transmission latency. The pre-processing latency refers to the time spent transforming input data to align with the requirements of the model. The model computational latency refers to the time it takes to perform the forward pass of a neural network, which involves feeding an input through the network, applying various mathematical operations, and producing an output. The post-processing latency refers to the time spent refining and interpreting the model's output after the model's forward pass. The transmission latency refers to the time it takes for data to travel from one tier to another in an Edge AI network.

The inference latency is collected via a timer that is started right before the launch of an inference test run and gets stopped when the model returns the output after successful execution. To that extent, we employ the default_timer from the timeit Python package for measuring inference latency. In the results of the five research questions, we employed the term "speedup" to signify the extent by which the median inference latency of a particular operator or tier is faster compared to the median inference latency of another operator or tier. The median inference latency here is the median value among the five repetitions, where each repetition involves running the inference test over 100 input data samples.

4.3.2 Accuracy

For different domain-specific models, the default accuracy metric varies. In our study, the employed metrics for evaluating the accuracy of image classification subjects like ResNet and ResNext are Top1% and Top5% accuracy, while the metric used for image segmentation subjects, like FCN and DUC, is mIoU (Mean Intersection Over Union). The definition of the metrics used is as follows:

- Top5% and Top1% accuracy: Top5% accuracy measures the proportion of validation samples where the true label is among the top 5 predictions with the highest confidence score. Top1% accuracy is a more strict evaluation metric, as it measures the proportion of validation samples where the model's highest confidence prediction matches the true label. Both Top1% and Top5% accuracy are useful metrics in image classification tasks and are often reported together to provide a more comprehensive evaluation of the model's performance. Therefore, in our study, we measure both metrics to gain a better understanding of how well image classification models are performing.
- mIoU%: This is a commonly used metric for evaluating the performance of image segmentation models. It measures the degree of overlap between the predicted segmentation masks and the ground truth masks and provides a measure of how well the model can accurately segment the objects in the image. The mIoU is calculated by first computing the Intersection over Union (IoU) for each class between the predicted mask and the ground truth mask, which is defined as the ratio of the intersection between the predicted and ground truth masks to their union. The IoU ranges from 0 to 1, with higher values indicating better overlap between the predicted and ground truth masks. The mIoU is then calculated as the average of the IoU scores across all classes in the dataset.

  The reason for using mIoU as an evaluation metric for image segmentation models is that it is sensitive to both false positives (areas predicted as belonging to a class when they do not) and false negatives (areas not predicted as belonging to a class when they should). This makes it a valuable metric for evaluating the overall accuracy of a segmentation model and can help identify areas where the model is performing poorly.

  In general, the accuracy metrics were calculated by validating each subject model on its specific validation data set, having varying sizes of images to get a more accurate and comprehensive picture of their accuracy performance, as shown in Table 3. The ResNet/ResNext subject models have been validated on the ILSVRC 2012 dataset (50k validation samples), while the FCN and DUC subject models were validated on the COCO 2017 dataset (5k validation samples) and CityScapes leftImg8bit dataset (500 validation samples), respectively. We used "accuracy" as a common term in the RQ results for the 4 subjects' respective accuracy metrics.

Controlled Variables The following variables are held constant across experiments to ensure internal validity and isolate the effects of the independent variables:

– **Model Architecture**: ResNe(x)t, FCN, and DUC.
– **Input Data**: Identical datasets and preprocessing are applied across all configurations.
– **Framework/Tools**: ONNX and Intel Neural Compressor.
– **Hardware Configuration**: CPU, RAM, and/or GPU of the respective Mobile, Edge, and Cloud tiers.

All five operator configurations are applied uniformly to the subject models listed in Table 3, ensuring a balanced experiment design in which each configuration is tested under equivalent conditions.

To ensure clarity regarding the experimental design across all research questions (RQs), we summarize the independent, dependent, and controlled variables in Table 4, titled *Summarization of Variables per Research Question*. This table outlines the factor levels and fixed conditions for each RQ, following guidelines from Wohlin et al. [112], and serves as a reference for understanding how the experimental variables are handled throughout the study.

Table 4: Summarization of Variables per Research Question

RQ	Independent Variables	Factor Levels [112]	Controlled Variables
RQ1	Operator, Tier, Bandwidth	1 op × 3 tier	Model Architecture, Input Data, Hardware Configs
RQ2	Operator, Tier, Bandwidth	1 op × 3 tier	Model Architecture, Input Data, Hardware Configs
RQ3	Operator, Tier, Bandwidth	1 op × 3 tier	Model Architecture, Input Data, Hardware Configs
RQ4	Operator, Tier, Bandwidth	1 op × 3 tier	Model Architecture, Input Data, Hardware Configs
RQ5	Operator, Tier, Bandwidth	2 op × 3 tier	Model Architecture, Input Data, Hardware Configs
RQ6	Operator, Tier, Bandwidth	5 op × 3 tier	Model Architecture, Input Data, Hardware Configs

4.4 Hypotheses

To formally test our research questions, we define statistical hypotheses involving the mean inference latency and accuracy across various deployment strategies. Below, we clarify the notation used in the hypothesis formulations:

– $\mu_{\text{latency}}(\text{Op}, \text{Tier})$ denotes the **mean inference latency** when operator Op is applied in deployment tier Tier.
– $\text{accuracy}(\text{Op}, \text{Tier})$ denotes the **inference accuracy** under the same conditions.
– Operator symbols (Op) refer to:
  – Op_{id}: identity (baseline operator)
  – Op_q: quantization
  – Op_e: early exit
  – Op_p: partitioning
  – Op_{qe}: quantized early exit (hybrid)
– Deployment tiers (Tier) include:
  – m, e, c: monolithic deployments on mobile, edge, and cloud
  – me, ec, mc: partitioned deployments across tiers
– BW_i represents the available network bandwidth (in Mbps), with $i \in \{1, 10, 50, 100, 150, 200\}$.

Each null hypothesis (H^0) assumes no statistically significant difference between the compared groups, while the alternative hypothesis (H^A) posits that at least one difference exists. The statistical hypotheses corresponding to our research questions are listed below.

RQ1: Is there a significant difference among the three Monolithic deployments in terms of inference latency?

$$\begin{aligned}
\mathrm{H}_{11}^{0}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_m) = \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_e) = \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_c) \\
\mathrm{H}_{11}^{A}: &\quad \exists\, i, j \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_j)
\end{aligned}$$

RQ1 (Monolithic Latency Comparison):

- H_{11}^{0}: There is no significant difference in latency across mobile, edge, and cloud for the identity operator.
- H_{11}^{A}: At least one pair of tiers has significantly different latency.

RQ2: Does the Quantization operator affect inference latency and accuracy within and across tiers?

$$\begin{aligned}
\mathrm{H}_{21}^{0}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_m) = \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_e) = \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_c) \\
\mathrm{H}_{21}^{A}: &\quad \exists\, i, j \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_j)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}_{22}^{0}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_k) = \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}_{22}^{A}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_k) \neq \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_k)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}_{23}^{0}: &\quad \text{accuracy}(\mathrm{Op}_{q}, \mathrm{Tier}_k) = \text{accuracy}(\mathrm{Op}_{id}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}_{23}^{A}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \text{accuracy}(\mathrm{Op}_{q}, \mathrm{Tier}_k) \neq \text{accuracy}(\mathrm{Op}_{id}, \mathrm{Tier}_k)
\end{aligned}$$

RQ2 (Quantization Effects):

- H_{21}^{0}: Quantization latency is equal across monolithic tiers.
- H_{21}^{A}: Latency differs across tiers for quantization.
- H_{22}^{0}: Quantization does not change latency compared to identity within each tier.
- H_{22}^{A}: Quantization changes latency compared to identity in at least one tier.
- H_{23}^{0}: Quantization does not change accuracy compared to identity in any tier.
- H_{23}^{A}: Accuracy is affected by quantization in at least one tier.

RQ3: Does the Early Exit operator impact inference latency and accuracy within and across tiers?

$$\begin{aligned}
\mathrm{H}_{31}^{0}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_m) = \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_e) = \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_c) \\
\mathrm{H}_{31}^{A}: &\quad \exists\, i, j \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_j)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}_{32}^{0}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_k) = \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}_{32}^{A}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{e}, \mathrm{Tier}_k) \neq \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_k)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}_{33}^{0}: &\quad \text{accuracy}(\mathrm{Op}_{e}, \mathrm{Tier}_k) = \text{accuracy}(\mathrm{Op}_{id}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}_{33}^{A}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \text{accuracy}(\mathrm{Op}_{e}, \mathrm{Tier}_k) \neq \text{accuracy}(\mathrm{Op}_{id}, \mathrm{Tier}_k)
\end{aligned}$$

RQ3 (Early Exit Effects):

- H^0_{31}: Early exit latency is equal across monolithic tiers.
- H^A_{31}: Latency differs across tiers for early exit.
- H^0_{32}: Early exit does not change latency compared to identity in any tier.
- H^A_{32}: Early exit affects latency in at least one tier.
- H^0_{33}: Early exit does not affect accuracy compared to identity.
- H^A_{33}: Accuracy changes due to early exit in at least one tier.

RQ4: Does the Partitioning operator affect inference latency across different deployment tiers?

$$\begin{aligned}
\mathrm{H}^0_{41}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_i) = \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_j) \quad \forall i, j \in \{me, ec, mc, m, e, c\} \\
\mathrm{H}^A_{41}: &\quad \exists\, i, j \in \{me, ec, mc, m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_{id}, \mathrm{Tier}_j)
\end{aligned}$$

RQ4 (Partitioning Effects):

- H^0_{41}: Latency is equal across all deployment types (monolithic and partitioned).
- H^A_{41}: At least one pair of deployment types shows a difference in latency.

RQ5: Do hybrid operators (e.g., Quantized Early Exit) impact inference latency and accuracy?

$$\begin{aligned}
\mathrm{H}^0_{51}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_m) = \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_e) = \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_c) \\
\mathrm{H}^A_{51}: &\quad \exists\, i, j \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_j)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}^0_{52}: &\quad \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_k) = \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}^A_{52}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_{qe}, \mathrm{Tier}_k) \neq \mu_{\text{latency}}(\mathrm{Op}_{q}, \mathrm{Tier}_k)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}^0_{53}: &\quad \text{accuracy}(\mathrm{Op}_{qe}, \mathrm{Tier}_k) = \text{accuracy}(\mathrm{Op}_{q}, \mathrm{Tier}_k) \quad \forall k \in \{m, e, c\} \\
\mathrm{H}^A_{53}: &\quad \exists\, k \in \{m, e, c\} \text{ such that } \text{accuracy}(\mathrm{Op}_{qe}, \mathrm{Tier}_k) \neq \text{accuracy}(\mathrm{Op}_{q}, \mathrm{Tier}_k)
\end{aligned}$$

RQ5 (Hybrid Operator Effects):

- H^0_{51}: Latency for the hybrid operator (quantized early exit) is equal across monolithic tiers.
- H^A_{51}: Latency differs across tiers for the hybrid operator.
- H^0_{52}: Latency of the hybrid operator matches that of quantization in each tier.
- H^A_{52}: Hybrid operator latency differs from quantization in at least one tier.
- H^0_{53}: Accuracy remains unchanged between hybrid and quantization operator.
- H^A_{53}: Accuracy differs between hybrid and quantization in at least one tier.

RQ6: What is the impact of bandwidth variations on the deployment strategies in terms of inference latency?

$$\begin{aligned}
\mathrm{H}^0_{61}: &\quad \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) = \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_j) \quad \forall i,j \in \{1,10,50,100,150,200\} \\
\mathrm{H}^A_{61}: &\quad \exists\, i,j \in \{1,10,50,100,150,200\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_j)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}^0_{62}: &\quad \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) = \mu_{\text{latency}}(\mathrm{Op}_y, \mathrm{Tier}_k, \mathrm{BW}_i) \quad \forall x,y \in \{\mathrm{id}, q, e, p, qe\} \\
\mathrm{H}^A_{62}: &\quad \exists\, x,y \in \{\mathrm{id}, q, e, p, qe\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_y, \mathrm{Tier}_k, \mathrm{BW}_i)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}^0_{63}: &\quad \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_l, \mathrm{BW}_i) \\
\mathrm{H}^A_{63}: &\quad \exists\, k,l \in \{m, e, c, me, mc, ec\} \text{ such that } \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_k, \mathrm{BW}_i) \neq \mu_{\text{latency}}(\mathrm{Op}_x, \mathrm{Tier}_l, \mathrm{BW}_i)
\end{aligned}$$

RQ6 (Bandwidth Sensitivity):

- H^0_{61}: Latency for a given operator and tier is invariant across bandwidth levels.
- H^A_{61}: Latency changes with bandwidth for at least one operator-tier combination.
- H^0_{62}: Different operators have similar latency under fixed bandwidth and tier.
- H^A_{62}: Operator choice affects latency under a given bandwidth-tier setting.
- H^0_{63}: Latency varies across tiers for a fixed operator and bandwidth.
- H^A_{63}: There is a tier-level effect on latency at fixed bandwidth and operator.

4.5 Study Design

The experiment follows a nested factorial design [112] in which some levels of one factor (e.g., deployment operator) are valid only for certain levels of another factor (e.g., deployment tiers). In our case: Partitioning and Quantized Early Exit Partitioned are nested within Multi-tiers (Mobile-Edge, Edge-Cloud, Mobile-Cloud) and Quantization, Early Exit, and Quantized Early Exit are nested within single tiers (Mobile, Edge, Cloud). Thus, specific operator-tier combinations are selected based on feasibility and/or practical relevance. This reduced testing scope makes the design a nested factorial design.

Table 5 summarizes the treatment combinations used in our study. For each research question (RQ), it outlines the selected deployment operators, tiers, and corresponding bandwidth settings, along with the sample size per treatment. This table provides a comprehensive view of how the nested factorial design was instantiated across different deployment strategies and experimental conditions.

Table 5: Summarization of the treatment combinations. Legend - M: Mobile, C: Cloud, E: Edge, I: Identity, Q: Quantized, E: Early Exit, P: Partition, QE: Quantized Early Exit, QEP: Quantized Early Exit Partition

RQ	Treatments (operators * tier * bandwidth)	Sample size per treatment
RQ1	Identity (I) * [M,E,C] * [ME: 200 Mbps, EC: 1 Mbps]	100 range of inputs
RQ2	Quantized (Q) * [M,E,C] * [ME: 200 Mbps, EC: 1 Mbps]	100 range of inputs
RQ3	Early Exit (E) * [M,E,C] * [ME: 200 Mbps, EC: 1 Mbps]	100 range of inputs
RQ4	Partition (P) * [ME,EC,MC] * [ME: 200 Mbps, EC: 1 Mbps]	100 range of inputs
RQ5	Quantized Early Exit (QE) * [M, E, C, ME, EC, MC] * [ME: 200 Mbps, EC: 1 Mbps] Quantized Early Exit Partition (QEP) * [M, E, C, ME, EC, MC] * [ME: 200 Mbps, EC: 1 Mbps]	100 range of inputs
RQ6	[I, Q, E, P, QE, QEP] * [M, E, C, ME, EC, MC] * [ME, EC: 1,10,50,100,150,200 Mbps]	1 single input

From the overall set of eight operator families for Edge AI inference discussed in Section 1, we narrow down the scope of our study to delve deeper into strategies that can optimize models in a black-box manner for deployment on resource-constrained and network-constrained Edge AI deployment scenarios. In other words, the operators transform the models as-is instead of fine-tuning models (which would invalidate prior model validation results). Among the model optimization operators, we selected three operators based on their representativeness (i.e., capable of addressing different aspects of model optimization, such as improving inference speed, providing better data privacy, optimizing resource usage, and reducing model size) and feasibility (i.e., the implementability of operators in black-box models).

Eventually, we selected one representative operator (i.e., Quantization) out of the three Model Compression operators (i.e., Quantization, Weight Pruning, Knowledge Distillation) as they all focus on a common goal of reducing the size and complexity of black-box models while preserving their accuracy as much as possible. Using the Quantization operator in the ONNX Runtime framework offered by the Intel Neural Compressor tool, Quantization can be performed by using the three widely used techniques discussed earlier, i.e., Static PTQ, Dynamic PTQ, and QAT. Our study used static PTQ as dynamic PTQ requires higher computational overhead during inference than static PTQ. QAT was excluded from the selection as it involved re-training models [46], while we focused on post-processing black-box models.

Model Partitioning was selected as it provides computational load splitting across the tiers (i.e., Mobile, Edge, and Cloud) of an Edge AI environment during distributed inference, enabling more efficient utilization of resources and providing scalable deployment of models. It also aims to provide better data privacy than the Monolithic Edge and Cloud deployments by transmitting intermediate outputs rather than the raw input data across the tiers of the Edge AI Environment.

The motivation for considering Early Exit as an operator is its aim to save computational resources and reduce the time required to predict by Exit early during the forward pass of the neural network. This would especially be valuable in scenarios where resources are constrained (i.e., Mobile and Edge tiers). For example, drones performing real-time object detection or navigation in constrained environments (e.g., disaster recovery or delivery scenarios) benefit from faster inferences through early Exit to make immediate decisions.

These three operators and their combinations are configured on black-box models for inference, depending on where the transformed (fragments of) black-box models will reside among the three tiers of the Edge AI Environment. This strategic configuration is vital for achieving optimal performance, minimizing latency, and improving the scalability of Edge AI deployment. By aligning these operators with the unique characteristics and constraints of each tier, a more effective and adaptable Edge AI ecosystem can be developed, catering to a wide range of use cases and scenarios.

We analyze the trade-off between two quantitative metrics, i.e., inference latency and inference accuracy. This analysis plays a crucial role in understanding the dynamic interplay between performance and latency within the Edge AI Environment, guiding the selection of optimal strategies based on a deployment Engineer's use cases and requirements. Some use cases might prioritize low latency at the expense of accuracy, while others could emphasize accuracy even if it leads to slightly higher latency [129]. The empirical data collected from deploying various operators on different tiers provides a quantitative basis for evaluating this trade-off and serves as a foundation for our long-term objective (outside the scope of this paper): the creation of recommendation systems to automatically suggest the most appropriate operators and deployment strategies for specific use cases, aligning with desired latency and accuracy requirements.

In our study, the deployment strategies are a mapping of deployment tiers to one or more deployment operators. The deployment tiers are the physical locations for model deployments, which include three Single-tier (i.e., Mobile, Edge, and Cloud) and three Multi-tier (i.e., Mobile-Edge, Edge-Cloud, Mobile-Cloud) environments. The Single-tiers refer to the deployment of entire models on single computing tiers to achieve Monolithic inference. The Multi-tiers refer to the deployment of Partitioned models across multiple computing tiers to achieve distributed inference. The deployment operators refer to the specific techniques that are applied to modify black-box models for efficient deployment and execution within the Edge AI Environment. They can be categorized into singular operators and hybrid operators. Singular operators are individual optimization techniques that are applied to models independently and hybrid operators are combinations of singular optimization techniques.

We considered four singular deployment operators, i.e., Identity Operator (no modifications), Quantization Operator, Early Exit Operator, and Partition Operator, and two hybrid operators, i.e., Quantized Early Exit and Quantized Early Exit Partitioned. We limited the hybrid operators to two as developing and evaluating hybrid operators involves combining multiple singular operators, which can increase the complexity of the study. By including a smaller set of hybrid operators, we can perform a more detailed comparative analysis against singular operators.

We perform Partitioning and Early Exit manually to check their feasibility (implementation), since automation across all types of models does not exist thus far. This feasibility analysis and the results of our study will help MLOps Engineers determine whether it is worth investing time/resources for

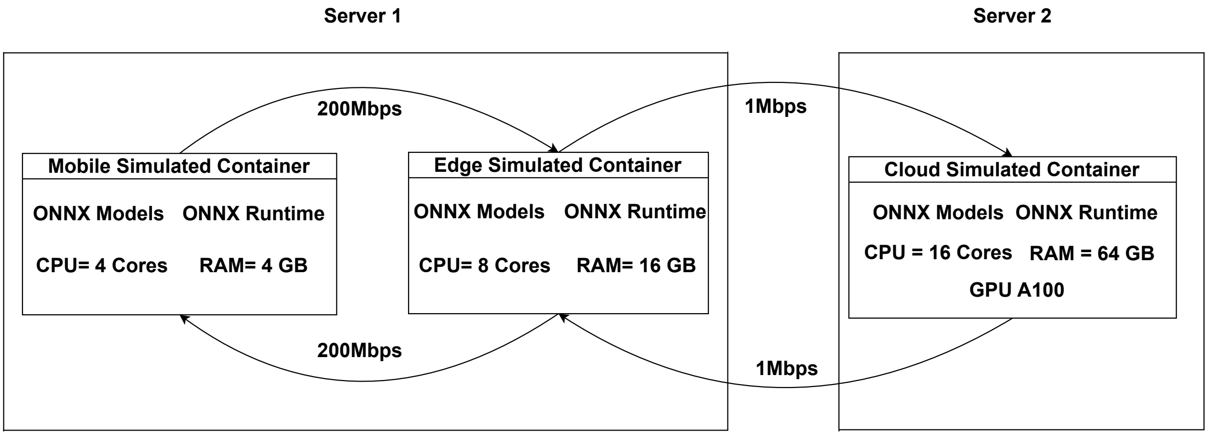


Fig. 2: Graphical illustration of Experimental Architecture for Edge AI

automating this in the future. For instance, the Early Exit criteria (i.e., skipping identically structured sub-graphs) require close analysis of ONNX computational graphs of the subject models, as explained in the Early Exit approach (Section 3.4). The Partitioning criteria (i.e., two equal-sized sub-models), require manually checking the sizes of sub-models, while the connection(s) used for achieving this criterion vary for subjects with varying architectures like ResNe(x)t, FCN, and DUC, as explained in our Partitioning approach (Section 3.2). Given the heterogeneity of the studied models, right now there is no automated tool that considers the mentioned criteria for these two operators across all models. To date, only one paper [69], uses a genetic algorithm to evenly split models into diverse sub-models, but this was only evaluated on the ResNet model of our study. As a consequence, we obtained Partitioning of Identity models, Early Exit of Identity/Quantized models, and Partitioning of Quantized Early Exit models through manual modifications. By executing these modified models in various deployment scenarios, across Mobile, Edge, and Cloud tiers, we then collect empirical data to analyze how each operator, and their combinations, perform under real-world conditions.

Previous studies [58,74] indicated that model Partitioning does not affect the inference accuracy, it just sends the same intermediate results remotely instead of within the same tier. As the model's architecture inherently involves sequential processing, the Partitioning aligns well with this logic by ensuring that data flows through each Partitioned model in a manner consistent with the original model's design, hence preserving accuracy. By default, the dimensions (width, height) of the model's input node of three subjects are fixed, i.e., (224, 224) for ResNet/ResNext and (800, 800) for DUC, while for one subject (i.e., FCN), it is dynamic (dependent on the width and height of input data). Therefore, for FCN, the intermediate data dimension/size varies for input data samples having varying widths/heights, while for other subjects, it remains fixed.

4.6 Experimental Execution

4.6.1 Experimental Setup

We simulated an Edge AI deployment architecture for Mobile, Edge, and Cloud tiers interconnected with each other as illustrated in Figure 2. This architecture is designed to support AI inference tasks in both a Monolithic and distributed manner for various deployment operators. Based on previous studies [93,92], we used Docker containers to simulate the hardware configurations of actual Mobile, Edge, and Cloud tiers. Docker is an open-source containerization technology ensuring a consistent and easily portable environment (or container) [78]. A docker container is a lightweight and portable package that includes all the necessary dependencies, libraries, and configurations to run a software application [78]. Docker provides a way to package and distribute applications in a standardized and portable format that can run on any platform, including Cloud, on-premise, and Edge [89].

This study utilized an experimental setup that may not fully generalize to all configuration scenarios. However, most configurations were selected based on prior studies to ensure comparability and validity.

Additionally, we experimented with independent variables that we considered the most impactful, such as network bandwidth and the selected models. We further discuss the impact of Docker containers, models, operators, and GPU-specific execution providers in detail in the Discussion and Threats to Validity sections (Section 6, Section 7). While both the Mobile and Edge tiers are simulated on a single server using Docker containers, resource constraints (e.g., CPU, RAM) and network conditions are carefully emulated to reflect the heterogeneous characteristics of these tiers. This setup enables a realistic evaluation of multi-tier deployment strategies in a controlled and repeatable environment. However, we acknowledge that certain hardware-specific characteristics, such as variations in physical device architectures, are not captured in this setup, which is common in similar simulation studies [93]. In particular, the Docker containers simulating Mobile and Edge tiers were configured with quad-core and octa-core CPUs (Intel(R) Xeon(R) E7-4870 2.40GHz) along with 4GB RAM and 16GB RAM, respectively. These configurations are based on previous studies [21,19]. We considered CPU-based Edge simulation to represent real-world scenarios where Edge devices often do not have dedicated GPUs due to power, size, or cost constraints. The Docker container simulating the Cloud was configured on a different server than the simulated Mobile and Edge containers. The simulated Cloud container contains 16-core CPUs (Intel(R) Xeon(R) Platinum 8268 CPU 2.90GHz), 64GB RAM, and an NVIDIA A100 GPU, these configurations are based on previous studies [53,94]. The Cloud runs all inference experiments on its GPU using ONNX Runtime with CUDA Execution Provider. Our simulated setup of Mobile, Edge, and Cloud Docker containers closely mimics real-world hardware configurations as mentioned below:

- The mobile Docker container mimics a lightweight Laptop (such as HP Chromebook x360) that has quad (4) core CPUs and 4 GB RAM [6].
- The edge container mimics a mini server (resource configuration of a server within a K8S Edge Cluster), which has 8 CPU cores and 16 GB RAM [19].
- The cloud container mimics a virtual machine with 16 cores, 64 GB RAM, and a GPU (Nvidia-A100 as this is the only available GPU in our lab server) [53,94].

For the three simulated Docker containers, we use the python:3.9-slim image as a base, on top of which we installed the necessary Python packages including a replica of the ONNX Run-time configuration (the out-of-the-box installation of the ONNX Run-time Python package). Here, the simulated Mobile/Edge/-Cloud device is a virtual representation of a physical Mobile/Edge/Cloud device created and operated within a software-based simulation environment (i.e., Docker). The Docker simulations provide a flexible and convenient way to configure and customize virtual environments that mimic various hardware specifications, network conditions, and software configurations of real-time deployment scenarios. Moreover, the advantages of cost-effectiveness and controlled testing make Docker simulations an invaluable tool for conducting inference experiments.

The simulated Mobile and Edge containers are interconnected to a common network bridge in Docker to exchange API requests with each other. The Edge container further connects with the external, simulated Cloud container. We configured the Linux Traffic Control utility [7] inside each configured Docker container for simulating Mobile-Edge and Edge-Cloud network bandwidths. After applying a given combination of operators, we placed the resulting .onnx files for the subject models on the corresponding simulated devices. The Flask Framework handles incoming and outgoing requests across Mobile, Edge, and Cloud devices. Base64 encoding is used while transferring data across the devices as it allows the data to be transmitted in a more reliable and universally readable format. For ResNet/ResNext and FCN/DUC models, the final output that is transmitted across the Edge AI Environment is the predicted label or the segmented image, respectively, which both have smaller sizes than the network bandwidths of the Edge AI Environment.

Similar to earlier work, [84,125,103,28,3], an Edge-Cloud bandwidth of 1 Mbps was used for simulating the WAN transmission latency, and the Mobile-Edge bandwidth of 200 Mbps was used for simulating the WLAN (Wireless Local Area Network) transmission latency. The selected bandwidth values aim to represent typical network conditions found in WAN and WLAN environments. WAN connections are prevalent for communicating between Edge and Cloud over large geographical distances and often have lower bandwidth due to factors like network congestion and long-distance transmission. On the other hand, WLAN connections commonly used for Mobile and Edge devices placed in closer proximity to

[6] https://www.hp.com/in-en/shop/hp-chromebook-x360-14a-ca0504tu-678m6pa.html
[7] https://man7.org/linux/man-pages/man8/tc.8.html

each other, tend to provide higher bandwidth. We also studied across a range of bandwidths, which is discussed in more detailed further below.

4.6.2 Operations done in preparation for each experiment

The entire validation set for running inference experiments is computationally expensive and time-consuming, especially when dealing with resource-constrained and network-constrained scenarios. Therefore, for analyzing the impact of input data on inference latency, we conducted the inference experiments for the subjects using a representative subset of 100 image samples selected from their specific validation sets with a specific criterion: we ensured that these image samples had larger sizes compared to the remaining validation set. Larger-sized images often present greater computational challenges due to increased memory requirements and processing complexity [52]. By selecting larger-sized image samples, the study can assess an upper bound for the inference latency performance and scalability of the models under investigation in resource-constrained and network-constrained scenarios. The recommendation for a minimum sample size of 100 is considered a typical number for the reliability of statistical analysis and to draw meaningful conclusions [37].

We conducted six inference latency trials to analyze the effects of varying deployment strategies, data sizes, and network bandwidths on inference latency across different scenarios. To ensure reliable and reproducible results, we divided the trials into two stages: an initial inference experiment and a final inference experiment. In the initial inference experiment, we performed 100 sequential runs on input test samples of varying sizes as a cache warm-up phase. This step stabilized the cache memory to mimic real-world continuous usage, where frequently accessed data populates the cache, as opposed to operating from a cold start. After the warm-up phase, we carried out the final inference experiment, which involved 500 sequential runs achieved through five repetitions on the same 100 input samples. These repetitions enhanced statistical significance and captured variability typical of real-world repetitive tasks. We logged the inference latency for each run in a text file and restarted the Mobile, Edge, and Cloud Docker containers after each experiment. A 20-second delay ensured consistent and isolated environments, further strengthening the reliability of our results.

To evaluate system performance under varying computational loads, we included models with different input sizes in our experiments. This consideration is critical for high-performance applications. Input size ranges across deployment strategies were as follows: ResNet/ResNeXt models ranged from 8 to 60 MB, DUC models from 19 to 22 MB, and FCN models from 2 to 5 MB.

Recognizing the correlation between model input size and network bandwidth and their impact on inference latency, we assessed the effect of Mobile-Edge and Edge-Cloud bandwidth variations. Based on commonly adopted practices in prior studies [1,17,84,125,130], we selected bandwidth values of 1, 10, 50, 100, 150, and 200 Mbps. To isolate the effect of network bandwidth and control for input size, we used the largest input sample in these experiments.

4.6.3 Measurement Procedure and Tools

As shown in Figure 3, our measurement infrastructure consists of 2 servers, i.e., We orchestrated the latency experiments as follows: Server 1 acted as the orchestrator, managing communication with the Cloud container on Server 2 (via the Mobile and Edge containers on Server 1), the Edge container on Server 1 (via the Mobile container on Server 2), and the Mobile container. Server 1 initiated the experiments and recorded raw latency data, storing the results in text files. We began each experiment with round-trip latency tests for different deployment strategies. To ensure consistency, we restarted the containers between subsequent deployment strategies and waited for 20 seconds to allow software reinitialization. During latency measurement, Server 1 sent input data to the Mobile container, which either processed the inference locally or offloaded computations to the Edge and Cloud containers. Once the containers completed the inference, they returned the output (predictions) to Server 1, completing the round trip.

We conducted inference latency experiments across a comprehensive set of configurations: twelve combinations of <Identity models(4), Monolithic tiers(3)>, twelve combinations of <Quantized models(4), Monolithic tiers(3)>, twelve combinations of <Early Exit models(4), Monolithic tiers(3)>, twelve combinations of <Partitioned models(4), Multi-tier setups(3)>, and twenty-four combinations of <Quantized Early Exit models(4), Deployment tiers(6)>. We evaluated model accuracy for Identity, Quantized, Early

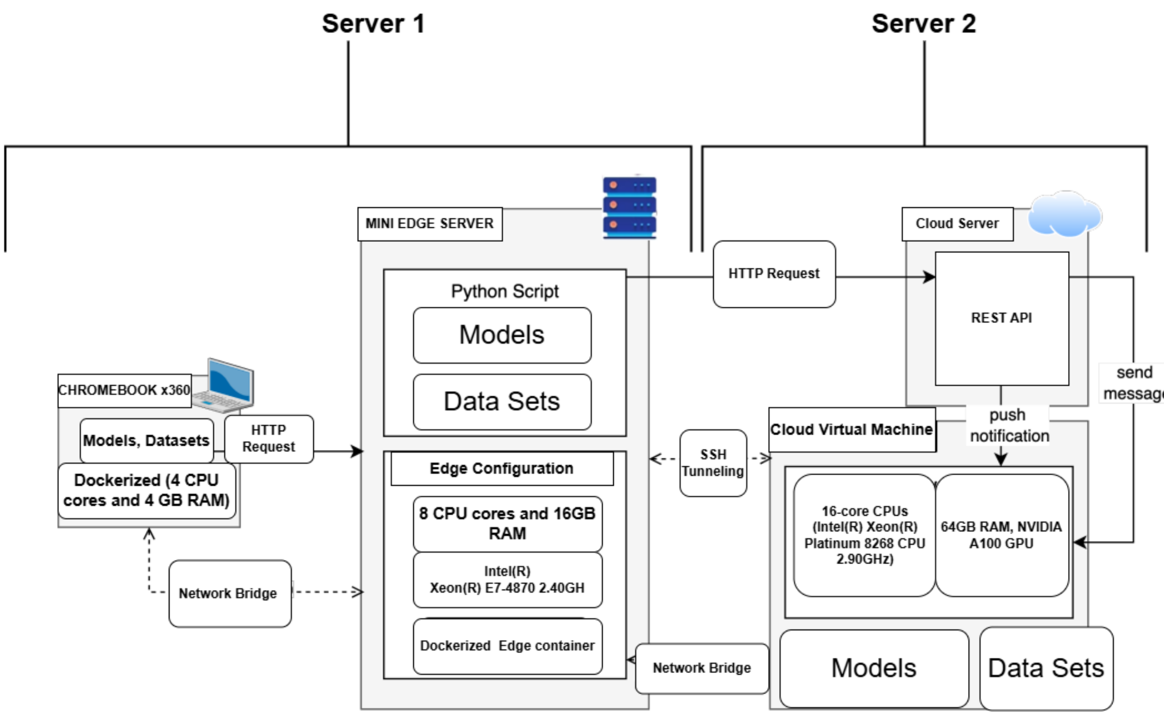


Fig. 3: Graphical illustration of Measurement Infrastructure

Exit, and Quantized Early Exit operators using their respective validation datasets, as shown in Table 9, and compared their accuracies.

We calculated the accuracy of the operators within each container using the complete validation dataset, as outlined in Figure 3. For Identity, Quantization, Early Exit, and Quantized Early Exit operators, we independently evaluated accuracy within CPU-based Docker environments (Mobile, Edge) and a GPU-based Docker environment (Cloud) to assess the impact of hardware on performance. This setup provided a holistic evaluation of model generalizability across diverse platforms. Both Mobile and Edge environments utilized the same Runtime Execution Provider (CPU) for inference and shared an Intel(R) Xeon(R) E7-4870 processor (2.40GHz). However, they differed in CPU and memory configurations, allowing us to perform a nuanced analysis of performance across varying hardware setups.

4.7 Data Analysis

The inference latency experiments' results are analyzed using various statistical methods. Firstly, we use the Shapiro-Wilks test and Q-Q plot for each deployment strategy to assess the normality of the inference latency distribution and determine if parametric or non-parametric tests are appropriate for testing the hypotheses. After observing from the Shapiro-Wilks test and Q-Q plot that the data does not conform to a normal distribution, we employ the Kruskal-Wallis (KW) test [87] to compare the inference latency distributions of independent groups (i.e., deployment strategies) based on two different dimensions (i.e., operator and tier dimension) and determine if there exists a significant difference among at least two of the independent groups (hypothesis testing). Further, the Conover test, a non-parametric post-hoc test, was used to perform pairwise comparisons of different deployment strategies across the tier and Operator Dimensions. This test was chosen to identify significant differences between strategies after a significant result was observed in the initial Kruskal-Wallis test. The Conover test's robustness to non-normal data ensures reliable comparisons of deployment strategies in scenarios where parametric assumptions may not hold. The design approach for the KW and Conover statistical tests across the tier and operator Dimension is illustrated in Table 6 and Table 7, respectively.

Table 6: Design of KW and Conover statistical tests for inference latency comparison across the tier(T) dimension, i.e., Mobile [M], Edge [E], Cloud [C], Mobile-Edge [ME], Edge-Cloud [EC], Mobile-Cloud [MC] for RQ1, RQ2, RQ3, RQ4, and RQ5.

T \ O	I	Q	E	P	QE	QEP
M	X					
E	X					
C	X					
ME						
EC						
MC						

(a) RQ1(X)

T \ O	I	Q	E	P	QE	QEP
M		X				
E		X				
C		X				
ME						
EC						
MC						

(b) RQ2(X)

T \ O	I	Q	E	P	QE	QEP
M			X			
E			X			
C			X			
ME						
EC						
MC						

(c) RQ3(X)

T \ O	I	Q	E	P	QE	QEP
M	X					
E	X					
C	X					
ME				X		
EC				X		
MC				X		

(d) RQ4(Y)

T \ O	I	Q	E	P	QE	QEP
M					X,Y	
E					X,Y	
C					X,Y	
ME						Y
EC						Y
MC						Y

(e) RQ5(X) RQ5(Y)

Table 7: Design of KW and Conover test statistical tests for inference latency comparison across the Operator(O) dimension, i.e., Identity (I), Quantized (Q), Early Exit (E), Partitioned (P), Quantized Early Exit (QE), and Quantized Early Exit Partitioned (QEP) for comparison across 6 operators (RQ2, RQ3, RQ4, RQ5)

T \ O	I	Q	E	P	QE	QEP
M	X	X	X		X	
E						
C						
ME				X		X

(a) Comparison across 6 operators (RQ2, RQ3, RQ4, RQ5)

T \ O	I	Q	E	P	QE	QEP
M						
E	X	X	X		X	
C						
ME				X		X

(b) Comparison across 6 operators (RQ2, RQ3, RQ4, RQ5)

T \ O	I	Q	E	P	QE	QEP
M						
E						
C	X	X	X		X	
ME				X		X

(c) Comparison across 6 operators (RQ2, RQ3, RQ4, RQ5)

4.7.1 Tier Dimension

As shown in Table 6, to investigate if there is a statistically significant difference between the three Monolithic deployment tiers (Single-tiers) in terms of inference latency of any of the Non-Partitioned operators, we perform a KW test ($\alpha = 0.05$) for each of the four Identity models in RQ1, four Quantized models in RQ2, four Early Exit models in RQ3, and four Quantized Early Exit models in RQ5 by comparing their latency across the three Single-tiers (Mobile, Edge, Cloud). These four variants of Identity, Quantized, Early Exit, and Quantized Early Exit models indicate for each of the subjects (i.e., ResNet, ResNext, FCN, and DUC). Moreover, to investigate if there is a statistically significant difference between the Monolithic deployment strategies and Multi-tier Partitioning strategies, for each subject, we conduct a KW test ($\alpha = 0.05$) across the Identity and Partitioned models in RQ4. We also perform this test across the Quantized Early Exit and Quantized Early Exit Partitioned models in RQ5, on their inference latency performance when deployed in the three Single-tier and three Multi-tier environments, respectively.

After observing significant differences (KW Test: p-value < 0.05), we further employ the Conover post-hoc test [15]. For the pairwise comparisons having significant differences (Conover test: adjusted p-value < 0.05), we evaluate Cliff's delta effect size [13] to assess the inference latency ranking of tiers based on the direction and magnitude of their difference in the corresponding RQs.

4.7.2 Operator Dimension

For each subject, to investigate if there is a statistically significant difference between the six operators (Identity, Quantized, Early Exit, Partitioned, Quantized Early Exit, Quantized Early Exit Partitioned), we perform the KW tests shown in Table 7. If there is a significant difference (KW Test: p-value $<$ 0.05), we further employ post-hoc Conover tests [15]. For the pairwise comparisons having significant differences (Conover test: adjusted p-value < 0.05), we used Cliff's delta effect size [13] to analyze how operators' inference latency ranks by looking at how much they differ (magnitude) and which way (sign) they differ in the corresponding RQs.

To evaluate the accuracy comparisons of the operators, we use the Wilcoxon Signed Rank Test [22] to determine if there exists a significant difference for the Identity vs Quantized models in RQ2, Identity vs Early Exit models in RQ3, Quantized vs Quantized Early Exit models in RQ4, and Early Exit vs Quantized Early Exit models in RQ5. This test analyzes the differences between two paired groups, i.e., the accuracy measurements under two different operators for the same subjects (i.e., ResNet, ResNext, FCN, DUC) and environments (i.e., Mobile, Edge, Cloud). Each group has 18 samples of accuracy measurements, i.e., six accuracy metric values ([ResNet, ResNext] x [Top 1%, Top 5%] + [FCN, DUC] x [mIOU%]) x 3 environments. In other words, in a particular operator's group, we are concatenating the accuracy metric(s), all of which are percentage values, of all four subjects, then comparing corresponding accuracy metrics using paired statistical tests.

To avoid false discoveries, Bonferroni Correction [42] is applied to the p-value for each Wilcoxon test comparing two operators by considering a p-value less than or equal to 0.0125 as statistically significant. The adjusted significance level of 0.0125 is derived by dividing the conventional significance level ($\alpha =$ 0.05) by the number of multiple comparisons (four in our case, as each operator is compared four times). After this correction, if a significant difference is observed between the two operators, we utilize Cliff's Delta effect size [13] to measure the magnitude/sign of their difference in corresponding RQs.

On the other hand, for the Shapiro-Wilks, Kruskal-Wallis, and Posthoc Conover tests, we compare the obtained p-value with the significance level of alpha = 0.05. According to [41], we interpret effect size as negligible ($d < 0.147$), small ($0.147 \leq d < 0.33$), medium ($0.33 \leq d < 0.474$), or large ($d \geq 0.474$). Negative values for d imply that, in general, samples from the distribution on the left member of the pair had lower values. These findings, in combination with the box plots displaying the distributions, will provide us with an additional understanding of the extent to which one deployment strategy differs from another. In particular, we analyze the median in the box plots to evaluate how each deployment strategy's performance compares to the others.

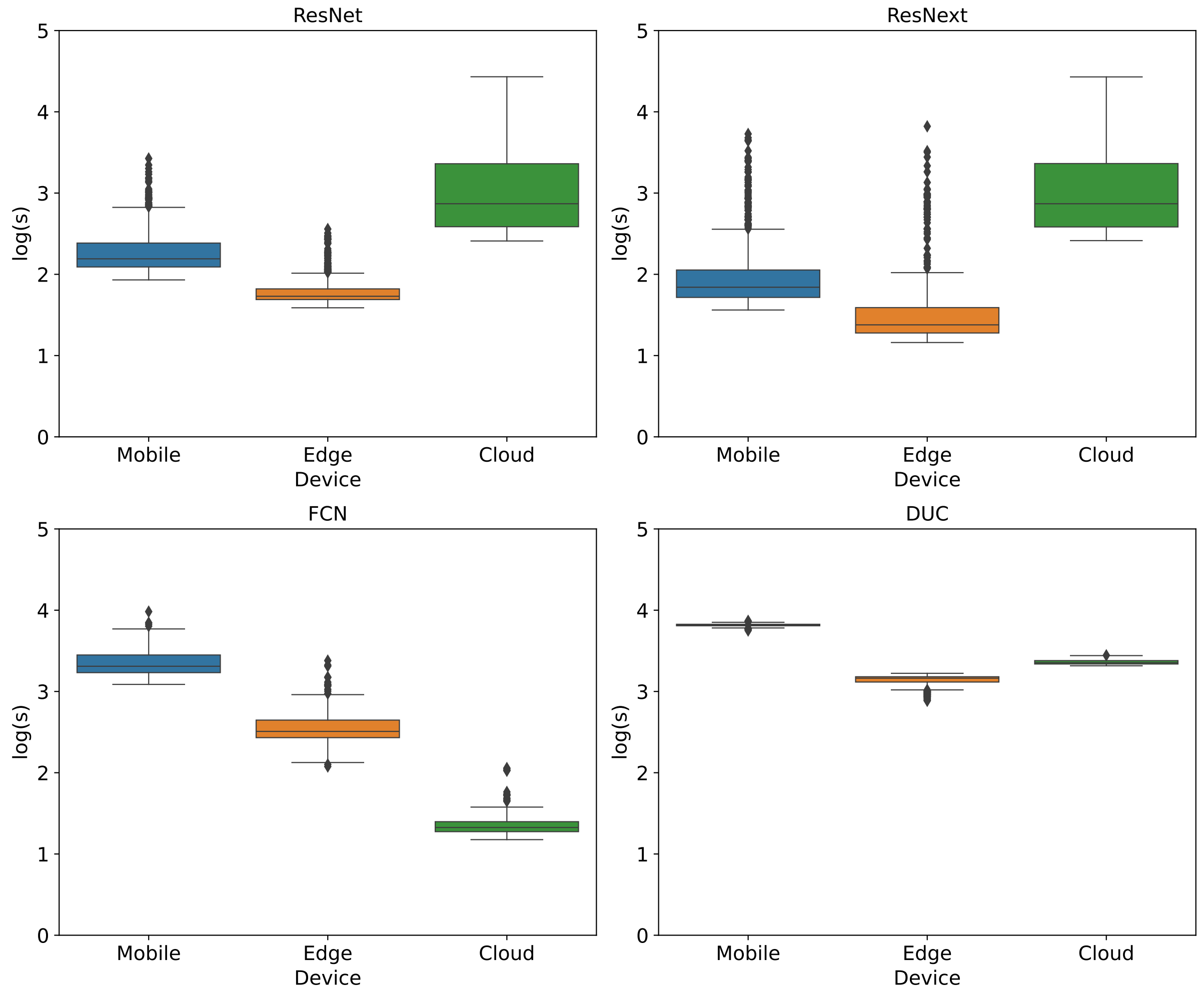


Fig. 4: Box plots of the measures collected for inference latency (seconds) from Mobile, Edge, and Cloud tiers for Identity versions of Subjects

5 Results

5.1 What is the impact of monolithic deployment in terms of inference latency and accuracy across the considered tiers? (RQ1)

5.1.1 Data Exploration

For the Identity models having large input data sizes (ResNet, ResNext, and DUC), the Edge tier shows the lowest median inference latency of 5.64, 3.96, and 23.64 seconds, respectively.

The box plots in Fig. 4 represent the inference latency distribution in the logarithmic scale of the subject's Identity models across the Mobile, Edge, and Cloud tiers. The Edge tier shows a 3.12x, 4.43x, and 1.21x lower median inference latency (Table 26) compared to the Cloud tier for ResNet, ResNext, and DUC Identity models, respectively, with a large effect size (Table 8). This is because of the Edge's higher network bandwidth capacity (200 Mbps) compared to Cloud (1 Mbps), which allows faster transmission of large input data samples for ResNet/ResNext (8 to 60 Mb) and DUC (19-22 Mb) during Edge deployment.

In contrast, the FCN Identity model shows the lowest median inference latency of 3.76 seconds in the Cloud tier, while the other 3 Identity models show the highest median inference latencies of 17.61, 17.61, and 28.61 seconds, respectively, in the Cloud tier (Table 26). This is because the input data samples for FCN are much smaller in size (2-5 Mb), which allows faster data transmission even in the network-constrained Cloud tier. Moreover, the Interquartile range (IQR) of the ResNet and ResNext is wider compared to FCN and DUC in the Cloud, indicating a more significant variability around the median. This is because the input data size range of ResNet and ResNext is around 17x higher compared to FCN and DUC, so a higher spread of inference latency values is observed for ResNet and ResNext. This behavior indicates that the input data sizes and network bandwidths play a critical role in the end-to-end inference latency during Edge and Cloud deployment.

The Edge tier shows 1.83x lower average median inference latency compared to the Mobile tier across the four Identity models. In particular, the Edge shows a drop of 3.30, 2.33, 15.10, and 21.81 seconds in median inference latency compared to Mobile for ResNet, ResNext, FCN, and DUC Identity models, respectively, along with a large effect size (Table 8). These differences are possibly due to the larger computational resources (CPU/RAM) of the Edge tier compared to the Mobile tier, due to which the computations during model inference are faster in the Edge tier. In particular, Edge has 4x the RAM and twice as many CPU cores as Mobile. The Cloud tier shows an 8.52 and 23.62 seconds drop in median inference latency compared to both Mobile and Edge tiers, respectively only for the FCN Identity model as the abundant computational resources of the Cloud (16x/4x RAM and 4x/2x CPU relative to Mobile/Edge) and the lowest input data size range for the FCN subject (as mentioned earlier) accounts for both faster model inference and data transmission.

The architectural complexity of models also plays a role in inference as among the 4 Identity models, the DUC Identity model shows the highest median inference latency of 45.46, 23.64, and 28.61 seconds on Mobile, Edge, and Cloud, respectively. The DUC Identity model has the highest number of graph nodes (i.e., 355), which contributes to the higher architectural complexity in comparison to the other Identity models (FCN: 260 graph nodes, ResNet/ResNext: 240 graph nodes). Here, the graph nodes represent the total number of operations in the ONNX computational graph of a model.

5.1.2 Normality Test

Normality assessment was conducted using the Shapiro-Wilk test and visual inspection through QQ plots. For each Identity model across all deployment tiers, the Shapiro-Wilk test resulted in $p < 0.05$, rejecting the null hypothesis of normality, as shown in Table 26. This suggests that inference latency distributions significantly deviate from normality.

Additionally, QQ plots showed heavy skewness and long tails across all subjects and tiers, as shown in Table 27, further supporting the result of the Shapiro-Wilk test. As a result, non-parametric methods were employed for statistical comparisons in the subsequent analysis.

5.1.3 Hypothesis Testing

Given the non-normal latency distributions, the Conover test (a post-hoc non-parametric multiple comparisons method) was used to evaluate significant differences between deployment tiers. Table 8 presents p-values and Cliff's Delta (δ) effect sizes for each pairwise comparison. All pairwise comparisons across Mobile (M), Edge (E), and Cloud (C) showed statistically significant differences ($p < 0.05$). The Cliff's Delta effect sizes were consistently large (L), indicating strong practical significance.

After rounding off the decimal digits (up to 4 places) in the accuracy metric values, we observed that for each subject, each of the 4 operators (i.e., Identity, Quantized, Early Exit, and Quantized Early Exit) exhibits identical performance between Mobile and Edge tiers, as shown in Table 9. The main reason seems to be the identical hardware (CPU processor) and software (packages) configuration of Mobile and Edge-simulated Docker containers. Table 9 shows a summary of results for RQ1.

Summary of Research Question 1

Among the three monolithic deployment tiers, the Edge tier could be the preferred choice in terms of latency in scenarios where models (ResNet, ResNext, DUC) have large input data size requirements and the Mobile/Cloud tier has computational/bandwidth limitations. In contrast, for models having smaller input data size requirements (FCN), Cloud deployment could be the optimal choice over computationally constrained tiers (Mobile/Edge).

Table 8: RQ1 results of the Cliff's Delta effect size and Conover test p-value between Mobile (M), Edge (E), and Cloud (C) deployment for I_t (Identity version of ResNet), I_x (Identity version of ResNext), I_f (Identity version of FCN), and I_d (Identity version of DUC).

I_x \ I_t	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	$9.6e^{-221}$	L (0.9)	$7.6e^{-187}$	-L (0.83)
E	$4.0e^{-99}$	L (0.71)	-	-	0.0	-L (1.0)
C	$2.9e^{-152}$	-L (0.86)	0	-L (0.92)	-	-

I_d \ I_f	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	0.0	L (0.99)	0.0	L (1.0)
E	0.0	L (1.0)	-	-	0.0	L (1.0)
C	0.0	L (1.0)	0.0	-L (1.0)	-	-

[1] In these, and later, tables, a Positive sign for the i^{th} cell shows that the latency of column[i]<row[i], while a Negative sign for the i^{th} cell shows that the latency of column[i]>row[i].
[2] In these, and later, tables, the L, M, S, and N symbols mean Large, Medium, Small, and Negligible effect size, respectively.
[3] In these, and later, tables, an empty cell means that the Cliff's Delta effect size was not considered because the pairwise comparison was not statistically significant based on the Conover test.

Table 9: Accuracy performance of Identity, Quantized, Early Exit, and Quantized Early Exit versions of subjects within Mobile, Edge, and Cloud tiers.

Subject	Operator	Model Size	Top-1%			Top-5%			mIOU%		
			Mobile	Edge	Cloud	Mobile	Edge	Cloud	Mobile	Edge	Cloud
ResNet	Identity	484 MB	82.52	82.520	82.522	96.008	96.008	96.008	-	-	-
ResNet	Quantized	123 MB	82.148	82.148	82.164	95.792	95.792	95.814	-	-	-
ResNet	Early Exit	380 MB	76.586	76.586	76.59	93.442	93.442	93.446	-	-	-
ResNet	Quantized Early Exit	96 MB	75.392	75.392	75.346	93.034	93.034	93.024	-	-	-
ResNext	Identity	319 MB	83.244	83.244	83.244	96.456	96.456	96.458	-	-	-
ResNext	Quantized	81 MB	83.084	83.084	83.14	96.402	96.402	96.386	-	-	-
ResNext	Early Exit	250 MB	77.276	77.276	77.284	93.92	93.92	93.924	-	-	-
ResNext	Quantized Early Exit	64 MB	75.668	75.668	75.616	93.83	93.83	93.826	-	-	-
FCN	Identity	199 MB	-	-	-	-	-	-	66.7343	66.7343	66.7348
FCN	Quantized	50 MB	-	-	-	-	-	-	66.38	66.38	66.35
FCN	Early Exit	164 MB	-	-	-	-	-	-	55.16	55.16	55.16
FCN	Quantized Early Exit	42 MB	-	-	-	-	-	-	54.36	54.36	54.32
DUC	Identity	249 MB	-	-	-	-	-	-	81.9220	81.9220	81.9223
DUC	Quantized	63 MB	-	-	-	-	-	-	81.62	81.62	81.62
DUC	Early Exit	215 MB	-	-	-	-	-	-	75.746	75.746	75.744
DUC	Quantized Early Exit	54 MB	-	-	-	-	-	-	75.32	75.32	75.32

Table 10: Summary of results for Research Question 1

Model	Best Tier	Latency (s)	Gain over Next Best	Notes
ResNet	Edge	5.64	3.12x faster than Cloud	Large input size, bandwidth bottleneck in Cloud
ResNext	Edge	3.96	4.43x faster than Cloud	Similar to ResNet
DUC	Edge	23.64	1.21x faster than Cloud	Highest graph complexity
FCN	Cloud	3.76	8.52s faster than Mobile	Small input size enables fast Cloud performance

5.2 What is the impact of the Quantization operator in terms of inference latency and accuracy within and across the considered tiers? (RQ2)

5.2.1 Data Exploration

In Mobile, the Quantized models show 1.17x higher and 1.27x lower average median inference latency w.r.t Identity models for ResNet/Resnext and FCN/DUC, respectively. In Edge, the Quantized models show 1.48x lower average median inference latency than the

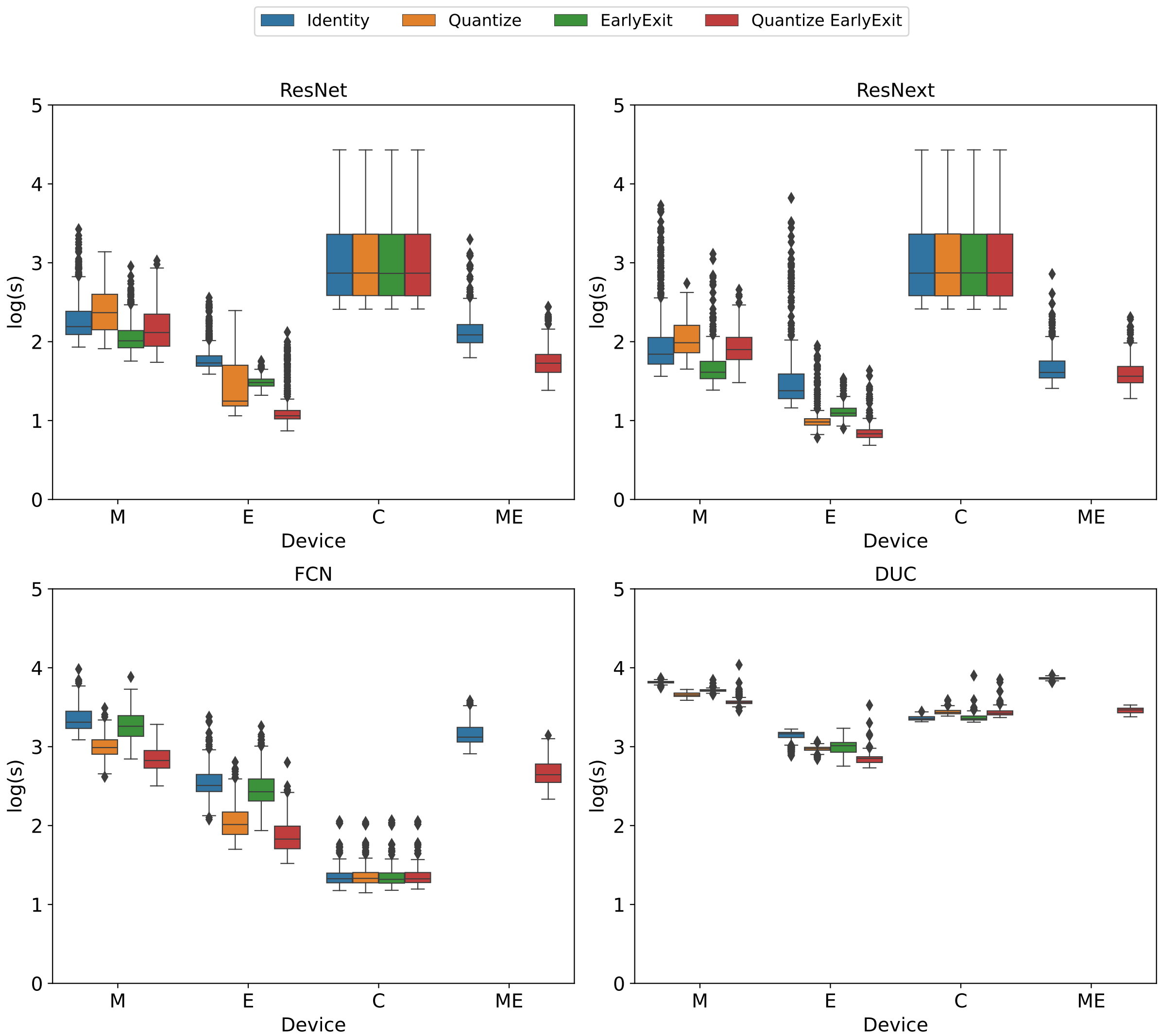


Fig. 5: Box plots of the measures collected for inference latency of Identity, Quantized, Early Exit, and Quantized Early Exit models in monolithic tiers (Mobile[M], Edge[E], Cloud[C]) and Partitioned Identity/Quantized Early Exit models in Mobile-Edge (ME) tier.

Identity models across all subjects. In Cloud, no significant difference was shown among Quantized and Identity models across all subjects (except DUC).

As shown in Figure 5 (blue and orange box plots), the Quantized models show 1.15x to 1.19x higher and 1.17x to 1.37x lower median inference latency compared to the Identity models (Table 27, Table 26) in the Mobile tier, along with small to large effect sizes (Table 11). The generated Quantized models are about 4x smaller than the Identity models of the subjects as shown in Table 9. Therefore, in an ideal situation, the Quantized models should show faster latency in comparison to Identity models due to model size reduction, as shown in FCN and DUC's latency results. But, for the ResNet and ResNext subjects, these models show slower latency, possibly due to their costly operations in the resource-constrained Mobile tier, leading to their higher CPU/Memory utilization than Identity models. In the Edge tier, the Quantized models show 1.20x to 1.64x lower median inference latency than the Identity models across the four subjects, along with large effect sizes (Table 12). This shows that the Quantized models perform faster than the Identity models in a high-resource environment (i.e., Edge).

During Cloud inference, the majority of graphical nodes (94.01% to 100%) are processed on the CUDA Execution provider for the Identity models, while for the Quantized models, 53.82% to 86.52% of the graphical nodes are processed on the CUDA Execution Provider and the remaining ones are processed on the CPU Execution Provider. Compared to Identity models, the percentage of graphical nodes that use CUDA Execution Provider is lower for Quantized models, which might imply that Quantized models are somewhat less optimized (or compatible) for GPU (i.e., CUDA) processing than Identity models. Among the four subjects, the Quantized model of DUC shows the lowest percentage of the graphical nodes (53.82%) placed on the CUDA Execution Provider, resulting in its significantly slower inference latency than the DUC Identity model, for which all the graphical nodes are placed on the CUDA execution provider. For the remaining three subjects (i.e., ResNet, ResNext, FCN), the Quantized models show similar inference latency compared to the Identity models in the Cloud tier due to the higher percentage of graphical nodes placed on the CUDA execution provider, i.e., 85.62% to 86.52% and 94.01% to 100% for Quantized and Identity models, respectively, which is statistically significant (Wilcoxon Test: p-value = 0.03, α = 0.05) along with large effect size (1.0).

The Quantized models show a small accuracy drop (<0.5%) across the four subjects in comparison to Identity models.

In terms of accuracy, the Quantized models demonstrate a marginal accuracy drop of 0.05% down to 0.38% compared to the Identity models, as shown in Table 9, While conducting the Wilcoxon test, the accuracy difference comes out to be statistically significant (p-value = 7.6e^{-6}, α = 0.0125). However, the effect size remains small (0.16). This trade-off suggests that Quantization is an effective operator for achieving faster inference in Edge (based on previous findings) without significantly compromising the accuracy performance.

When comparing Quantization across the three monolithic deployment tiers, the Quantized models show 2.60x and 3.44x lower average median inference latency during Edge deployment than their deployment in Mobile and Cloud, respectively.

The Quantized models across all subjects in the Edge exhibit 1.97x to 3.06x lower median inference latency compared to their inference in the Mobile, as shown in Figure 5 (orange box plots), along with large effect sizes (Table 14) due to its higher computational resources. For ResNet, ResNext, and DUC subjects, the Quantized models in the Cloud demonstrate 1.58x to 6.61x higher median inference latency than their inference in the Edge tier due to the higher impact of the large input data sizes of these subjects on the transmission across the restricted Edge-Cloud network. However, the Quantized model for the FCN subject exhibits 1.97x lower median inference latency in the Cloud tier compared to the Edge tier with a large effect size (Table 14). This is due to the FCN's small input data sizes leading to a lower impact on data transmission across the restricted Edge-Cloud network. The reasoning behind these findings is similar and explained briefly in the RQ1 findings (Section 5.1). Table 14 shows a summary of results for RQ2.

5.2.2 Normality Test

The Shapiro-Wilk test and QQ plots were applied to assess the normality of latency distributions. For most model-subject combinations, the null hypothesis of normality was rejected ($p < 0.05$), indicating non-normal distributions of inference latency values(see Table 27). As a result, non-parametric methods like the Conover and Wilcoxon tests were selected for hypothesis testing.

5.2.3 Hypothesis Testing

According to the Conover test, the null hypothesis that there is no significant inference latency difference between the Quantized and Identity models was rejected for the DUC subject (2.0e^{-247}, α = 0.05) during Cloud deployment. The large effect size (Table 13) for DUC shows that its Quantized model has a significantly higher distribution of inference latency magnitude compared to its Identity model. Conversely, for the remaining three subjects (i.e., ResNet, ResNext, FCN), the null hypothesis cannot be rejected, indicating that their Quantized models show similar or equivalent inference latency compared to their Identity models.

Table 11: Cliff's Delta effect size and Conover Test p-value between I_{M}, Q_{M}, E_{M}, P_{ME}, QE_{M}, and QEP_{ME} versions of subjects for RQ2, RQ3, RQ4, and RQ5

R_{t} / R_{x}	I_{M}		Q_{M}		E_{M}		P_{ME}		QE_{M}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{M}	-	-	$4.0e^{-9}$	-S (0.24)	$4.4e^{-67}$	L (0.58)	$2.8e^{-26}$	M (0.37)	$3.0e^{-20}$	S (0.26)	$3.9e^{-270}$	L (0.94)
Q_{M}	$2.2e^{-23}$	-M (0.34)	-	-	$1.7e^{-113}$	L (0.69)	$2.6e^{-59}$	L (0.54)	$3.2e^{-50}$	M (0.44)	0.0	L (0.96)
E_{M}	$9.1e^{-88}$	L (0.6)	$5.8e^{-179}$	L (0.79)	-	-	$2.3e^{-12}$	-S (0.25)	$4.0e^{-17}$	-S (0.24)	$5.3e^{-94}$	L (0.80)
P_{ME}	$7.0e^{-94}$	L (0.63)	$8.3e^{-187}$	L (0.83)			-	-			$6.5e^{-157}$	L (0.87)
QE_{M}	0.001	-N (0.12)	$6.6e^{-12}$	S (0.29)	$7.8e^{-114}$	-L (0.68)	$1.3e^{-120}$	-L (0.71)	-	-	$6.2e^{-171}$	L (0.85)
QEP_{ME}	$2.8e^{-148}$	L (0.75)	$9.8e^{-253}$	L (0.91)	$5.2e^{-12}$	S (0.24)	$8.3e^{-10}$	S (0.23)	$2.5e^{-179}$	L (0.82)	-	-

FCN / DUC	I_{M}		Q_{M}		E_{M}		P_{ME}		QE_{M}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{M}	-	-	$8.1e^{-292}$	L (0.92)	$6.8e^{-18}$	S (0.25)	$6.2e^{-92}$	L (0.65)	0.0	L (0.99)	0.0	L (1.0)
Q_{M}	0.0	L (1.0)	-	-	$1.0e^{-196}$	-L (0.77)	$5.3e^{-83}$	-L (0.58)	$2.2e^{-64}$	L (0.56)	$2.0e^{-191}$	L (0.88)
E_{M}	0.0	L (0.99)	$7.7e^{-322}$	-L (0.93)	-	-	$2.0e^{-34}$	M (0.38)	0.0	L (0.94)	0.0	L (0.99)
P_{ME}	0.0	-L (0.98)	0.0	-L (1.0)	0.0	-L (1.0)	-	-	$3.7e^{-250}$	L (0.88)	0.0	L (0.99)
QE_{M}	0.0	L (0.99)	$1.2e^{-321}$	L (0.94)	0.0	L (0.98)	0.0	L (1.0)	-	-	$9.5e^{-46}$	L (0.57)
QEP_{ME}	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)	0.0	L (0.99)	-	-

[1] I_{M}, Q_{M}, E_{M}, QE_{M} denote Identity, Quantized, Early Exit, and Quantized Early Exit models in the Mobile tier.
[2] P_{ME} and QEP_{ME} denote Partitioned and Quantized Early Exit Partitioned models in the Mobile-Edge tier.
[3] L, M, S, and N refer to Large, Medium, Small, and Negligible effect sizes, respectively.
[4] An empty p cell indicates the pairwise comparison was not statistically significant based on the Conover test.

Table 12: Cliff's Delta effect size and Conover Test p-value between I_{E}, Q_{E}, E_{E}, P_{ME}, QE_{E}, and QEP_{ME} versions of subjects for RQ2, RQ3, RQ4, and RQ5

R_{t} / R_{x}	I_{E}		Q_{E}		E_{E}		P_{ME}		QE_{E}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{E}	-	-	$3.8e^{-147}$	L (0.58)	$9.3e^{-161}$	L (0.99)	$2.9e^{-103}$	-L (0.83)	0.0	L (0.91)	0.0003	N (0.13)
Q_{E}	0.0	L (0.9)	-	-			0.0	-L (0.83)	$1.0e^{-105}$	L (0.74)	$2.7e^{-114}$	-L (0.55)
E_{E}	$3.6e^{-194}$	L (0.93)	$7.5e^{-44}$	-L (0.69)	-	-	0.0	-L (1.0)	$4.5e^{-94}$	L (0.81)	$8.8e^{-127}$	-L (0.90)
P_{ME}	$1.9e^{-74}$	-L (0.53)	0.0	-L (0.95)	0.0	-L (1.0)	-	-	0.0	L (0.99)	$5.1e^{-135}$	L (0.87)
QE_{E}	0.0	L (0.98)	$5.3e^{-86}$	L (0.78)	$4.4e^{-219}$	L (0.93)	0.0	L (1.0)	-	-	0.0	-L (0.90)
QEP_{ME}	$4.6e^{-36}$	-M (0.42)	0.0	-L (0.93)	0.0	-L (0.99)	$1.4e^{-9}$	S (0.23)	0.0	-L (0.99)	-	-

FCN / DUC	I_{E}		Q_{E}		E_{E}		P_{ME}		QE_{E}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{E}	-	-	$1.3e^{-302}$	L (0.92)	$7.7e^{-14}$	S (0.25)	0.0	-L (0.98)	0.0	L (0.98)	$6.2e^{-46}$	-L (0.48)
Q_{E}	$7.9e^{-311}$	L (0.90)	-	-	$5.9e^{-219}$	-L (0.83)	0.0	-L (1.0)	$3.0e^{-36}$	L (0.5)	0.0	-L (0.97)
E_{E}	$3.9e^{-239}$	L (0.83)	$2.2e^{-10}$	-S (0.25)	-	-	0.0	-L (0.98)	0.0	L (0.95)	$2.1e^{-99}$	-L (0.57)
P_{ME}	0.0	-L (1.0)	0.0	-L (1.0)	0.0	-L (1.0)	-	-	0.0	L (1.0)	$1.4e^{-185}$	L (0.99)
QE_{E}	0.0	L (0.99)	$4.8e^{-249}$	L (0.94)	$3.1e^{-321}$	L (0.82)	0.0	L (1.0)	-	-	0.0	-L (1.0)
QEP_{ME}	$6.4e^{-207}$	-L (1.0)	0.0	-L (1.0)	0.0	-L (1.0)	$7.3e^{-166}$	L (1.0)	0.0	-L (1.0)	-	-

[1] I_{E}, Q_{E}, E_{E}, QE_{E} denote Identity, Quantized, Early Exit, and Quantized Early Exit models in the Edge tier.
[2] P_{ME} and QEP_{ME} denote Partitioned and Quantized Early Exit Partitioned models in the Mobile-Edge tier.
[3] L, M, S, and N refer to Large, Medium, Small, and Negligible effect sizes, respectively.
[4] An empty p cell indicates the pairwise comparison was not statistically significant based on the Conover test.

Summary of Research Question 2

The Quantization operator could be the preferred choice over the Identity operator across the four subjects when faster latency (1.48x) is a concern in the Edge tier, at a small accuracy loss (<0.4%). Among the three monolithic deployment tiers, the Edge again is the most suitable deployment tier for the Quantization operator when factors like large input data size and constrained computational (Mobile)/ network (Cloud) environment play a crucial role. In contrast, Cloud deployment again is a better option for this operator when factors like small input data sizes and constrained computational environments (Mobile/Edge) are important.

Table 13: Cliff's Delta effect size and Conover Test p-value between I_{C}, Q_{C}, E_{C}, P_{ME}, QE_{C}, and QEP_{ME} versions of subjects for RQ2, RQ3, RQ4, and RQ5

ResNext \ ResNet	I_{C}		Q_{C}		E_{C}		P_{ME}		QE_{C}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{C}	-	-					$4.6e^{-257}$	L (0.95)			0.0	L (1.0)
Q_{C}			-	-			$6.6e^{-257}$	L (0.95)			0.0	L (1.0)
E_{C}					-	-	$3.6e^{-257}$	L (0.95)			0.0	L (1.0)
P_{ME}	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)	-	-	$4.6e^{-257}$	-L (0.95)	$1.3e^{-50}$	L (0.87)
QE_{C}							0.0	-L (1.0)	-	-	0.0	L (1.0)
QEP_{ME}	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)	0.0001	S (0.23)	0.0	L (1.0)	-	-

DUC \ FCN	I_{C}		Q_{C}		E_{C}		P_{ME}		QE_{C}		QEP_{ME}	
	p	δ	p	δ	p	δ	p	δ	p	δ	p	δ
I_{C}	-	-					0.0	-L (1.0)			$8.0e^{-298}$	-L (1.0)
Q_{C}	$1.9e^{-289}$	-L (0.92)	-	-			0.0	-L (1.0)			$1.9e^{-287}$	-L (1.0)
E_{C}	0.02	-N (0.03)	$1.4e^{-263}$	L (0.86)	-	-	0.0	-L (1.0)			$5.5e^{-301}$	-L (1.0)
P_{ME}	0.0	-L (1.0)	$3.1e^{-276}$	-L (1.0)	0.0	-L (1.0)	-	-	0.0	L (1.0)	$2.6e^{-57}$	L (0.99)
QE_{C}	$8.9e^{-221}$	-L (0.85)	$6.8e^{-10}$	S (0.20)	$1.5e^{-196}$	-L (0.78)	0.0	L (1.0)	-	-	$5.3e^{-296}$	-L (1.0)
QEP_{ME}	0.0	-L (0.94)	$5.3e^{-16}$	-S (0.30)	0.0	-L (0.90)	$6.3e^{-188}$	L (1.0)	$3.8e^{-45}$	-M (0.42)	-	-

1 I_{C}, Q_{C}, E_{C}, QE_{C} denote Identity, Quantized, Early Exit, and Quantized Early Exit models in the Cloud tier.
2 P_{ME} and QEP_{ME} denote Partitioned and Quantized Early Exit Partitioned models in the Mobile-Edge tier.
3 L, M, S, and N refer to Large, Medium, Small, and Negligible effect sizes, respectively.
4 An empty p cell indicates the pairwise comparison was not statistically significant based on the Conover test.

Table 14: RQ2 results of the Cliff's Delta effect size and Conover test p-value between Mobile (M), Edge (E), and Cloud (C) deployment of Quantized models (Q_{t}, Q_{x}, Q_{f}, Q_{d} denote Quantized versions of ResNet, ResNext, FCN, and DUC respectively).

Q_{x} \ Q_{t}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	$3.3e^{-230}$	L (0.93)	$6.5e^{-142}$	-L (0.74)
E	0.0	L (0.99)	-	-	0.0	-L (1.0)
C	0.0	-L (0.98)	0.0	-L (1.0)	-	-

Q_{d} \ Q_{f}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	0.0	L (1.0)	0.0	L (1.0)
E	0.0	L (1.0)	-	-	0.0	L (0.99)
C	0.0	L (1.0)	0.0	-L (1.0)	-	-

1 A Positive sign for the i^{th} cell shows that the latency of column[i] < row[i], while a Negative sign means column[i] > row[i].
2 L, M, S, and N symbols denote Large, Medium, Small, and Negligible effect sizes, respectively.
3 Empty cells indicate non-significant comparisons based on the Conover test.

Model	Best Tier	Latency Change	Accuracy Drop	Notes
ResNet	Edge	1.64x faster than Identity	0.37%	Costly on Mobile; Cloud uses more CPU than GPU
ResNext	Edge	1.59x faster than Identity	0.16%	Same as ResNet; best in Edge, neutral in Cloud
FCN	Cloud	1.97x faster than Edge	0.35%	Small input size favors Cloud; ideal for low-bandwidth
DUC	Edge	1.58x faster than Cloud	0.30%	Cloud performance bottlenecked by GPU compatibility

Table 15: Summary of results for Research Question 2

5.3 What is the impact of the Early Exit operator in terms of inference latency and accuracy within and across the considered tiers? (RQ3)

5.3.1 Data Exploration

In Mobile and Edge, the Early Exit models show a 1.15x and 1.21x lower average median inference latency than the Identity models, respectively. In the Cloud, the Early Exit mod-

els show no practically significant difference in inference latency compared to the Identity models.

The Early Exit models show 1.05x to 1.25x and 1.08x to 1.32x lower median inference latency than the Identity models (Table 28, Table 26) in the Mobile and Edge tiers, respectively, as shown in Figure 5 (blue and green box plots), along with small or large effect sizes (Table 11, Table 12). The utilization of intermediate predictions in the Early Exit models allows it to leverage information from earlier stages of the neural network, leading to model size reduction (1.15x to 1.27x) and faster inference compared to the Identity models in restricted-constrained tiers (i.e., Mobile and Edge).

In Mobile, the Quantized models show 1.44x higher and 1.18x lower average median inference latency w.r.t Early Exit models for ResNet/ResNext and FCN/DUC, respectively. In Edge, the Quantized models show 1.23x lower average median inference latency than the Early Exit models across the four subjects. In the Cloud, no significant inference latency difference was shown among Quantized and Early Exit models across all subjects (except DUC).

During Mobile deployment, the Quantized models for ResNet and ResNext subjects show 1.42x to 1.45x higher median inference latency than the Early Exit models, as shown in Figure 5 (orange and green box-plots), along with large effect sizes (Table 11). In contrast, for FCN and DUC subjects, the Quantized models show 1.05x to 1.31x lower median inference latency than the Early Exit models, along with large effect sizes (Table 11). Even though the Quantized model sizes of the ResNet and ResNext subjects are 3.08x lower than the Early Exit models, they still show slower latency results, possibly due to the costly Quantization operations of these two subjects in a low-resource environment (i.e., Mobile) similar to the reasoning discussed in RQ1 (Section 5.1) when comparing Quantized models with Identity models in the Mobile tier.

In the Edge tier, the Quantized models show 1.03x to 1.51x lower median inference latency than the Quantized models across the four subjects. In comparison to Early Exit models (Table 12), the inference latency of Quantized models is similar (ResNet) or significantly faster with small (DUC) or large (ResNext, FCN) effect sizes. This shows that in the Edge tier, the Quantized models are overall a better option than the Early Exit models in terms of faster latency.

The Early Exit models cost a medium accuracy drop of 2.53% to 11.56% and 2.35% to 11.21% in comparison to Identity and Quantized models, respectively.

In terms of accuracy, the Early Exit models show a significant statistical difference (Wilcoxon Test: p-value = 7.6e^{-6}, $\alpha = 0.0125$) with a medium effect size (0.38) in comparison to Identity and Quantized models. As presented in Table 9, the Early Exit models reveal an accuracy drop ranging from 2.53% to 11.56% and 2.35% to 11.21% when compared with Identity and Quantized models. This finding suggests that Early Exit models are less effective in accuracy performance relative to both Identity and Quantized models. Based on this and previous findings, the Quantization operator can achieve faster inference while maintaining a reasonably high level of accuracy, making it a preferred choice over the early exit operator.

When comparing Early Exiting across the three monolithic deployment tiers, the Early Exit models show 1.92x and 2.90x lower average median inference latency during Edge deployment than their deployment in Mobile and Cloud tiers, respectively.

The Early Exit models in the Edge outperform the ones on the Mobile tier by 1.69x to 2.29x in terms of median inference latency, as shown in Figure 5 (green box plots), with large effect sizes (Table 16) due to the higher computational resources of Edge. For ResNet, ResNext, and DUC subjects, the Early Exit models in the Cloud tier exhibit 1.40x to 5.91x higher median inference latency compared to the Edge, with large effect sizes (Table 16) due to the higher impact of the large input data sizes of these subjects on the transmission across the restricted Edge-Cloud network during Cloud deployment. However, for the FCN subject, the Early Exit model in the Cloud tier experiences a 3.03x lower median inference latency compared to the Edge tier, with a large effect size (Table 16). This is due to the FCN's small input data sizes leading to a lower impact on data transmission across the restricted Edge-Cloud network during Cloud deployment. The reasoning behind these findings is similar and explained briefly in the RQ1 findings (Section 5.1). Table 16 shows a summary of results for RQ3.

5.3.2 Normality Test

The Shapiro-Wilk test and QQ plots reveal non-normality across all configurations and deployment tiers for latency data (Table 28), justifying the use of non-parametric tests such as the Conover test and Cliff's Delta for pairwise comparisons.

5.3.3 Hypothesis Testing

According to the Conover test, for ResNet, ResNext, and FCN, the null hypothesis that there is no significant difference between the Early Exit and Identity model in the Cloud cannot be rejected, indicating that the Early Exit models during Cloud deployment show similar or equivalent inference latency compared to the Identity models. For the DUC subject, the null hypothesis was rejected, although with a negligible effect size (Table 13), suggesting that the difference is likely not practically significant. The main reason for not having a significant difference is the ample availability of computational resources in the Cloud tier compared to resource-constrained Mobile and Edge tiers due to which the impact of Early Exiting on subjects is not significant in comparison to the Identity models in the Cloud tier.

In the Cloud tier, the Quantized models show no statistically significant difference (according to the Conover test) in inference latency in comparison to Early Exit models for all subjects (except DUC). The reasoning for these findings is due to the lower compatibility of DUC's Quantization nodes with the CUDA Execution Provider during Cloud deployment, which is similar to those discussed briefly in the first finding of RQ1 (Section 5.1) when comparing Quantized and Identity models in the Cloud tier.

Table 16: RQ3 results of the Cliff's Delta effect size and Conover test p-value between Mobile (M), Edge (E), and Cloud (C) deployment of Early Exit models (E_{t}, E_{x}, E_{f}, and E_{d} denote Early Exit versions of ResNet, ResNext, FCN, and DUC respectively).

E_{t} \ E_{x}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	0.0	L (1.0)	0.0	-L (0.98)
E	0.0	L (0.9)	-	-	0.0	-L (1.0)
C	0.0	-L (0.98)	0.0	-L (1.0)	-	-

E_{f} \ E_{d}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	0.0	L (0.9)	0.0	L (1.0)
E	0.0	L (1.0)	-	-	0.0	L (1.0)
C	0.0	L (1.0)	0.0	-L (1.0)	-	-

[1] A Positive sign for the i^{th} cell shows that the latency of column[i] < row[i], while a Negative sign means column[i] > row[i].
[2] L, M, S, and N symbols denote Large, Medium, Small, and Negligible effect sizes, respectively.
[3] Empty cells indicate non-significant comparisons based on the Conover test.

Summary of Research Question 3

Similar to RQ2, the Quantized operator could be the preferred choice when faster latency (1.23x) is a concern in the Edge tier, at medium accuracy improvement (up to 11.21%) than the Early Exit operator, which shows faster latency (1.21x) than the Identity models at medium accuracy drop (up to 11.56%). Among the three monolithic deployment tiers, the Edge again is the most suitable deployment tier for the Early Exit operator when factors like large input data size and a constrained network(Cloud) environment play a crucial role. Cloud deployment is again a better option for this operator when factors like small input data size and constrained computational environments (Mobile/Edge) play a crucial role.

Model	Best Tier	Latency Gain	Accuracy Drop	Notes
ResNet	Edge	1.21x faster than Identity	11.56%	No significant gain in Cloud; benefits from intermediate prediction on Edge
ResNext	Edge	1.21x faster than Identity	10.77%	Similar to ResNet; best used in constrained environments
FCN	Cloud	No significant gain	2.53%	Cloud benefits due to small input size; no Edge advantage
DUC	Edge	1.08x–1.32x faster	6.80%	Gains seen in Edge; Cloud gain negligible due to GPU usage limits

Table 17: Summary of results for Research Question 3

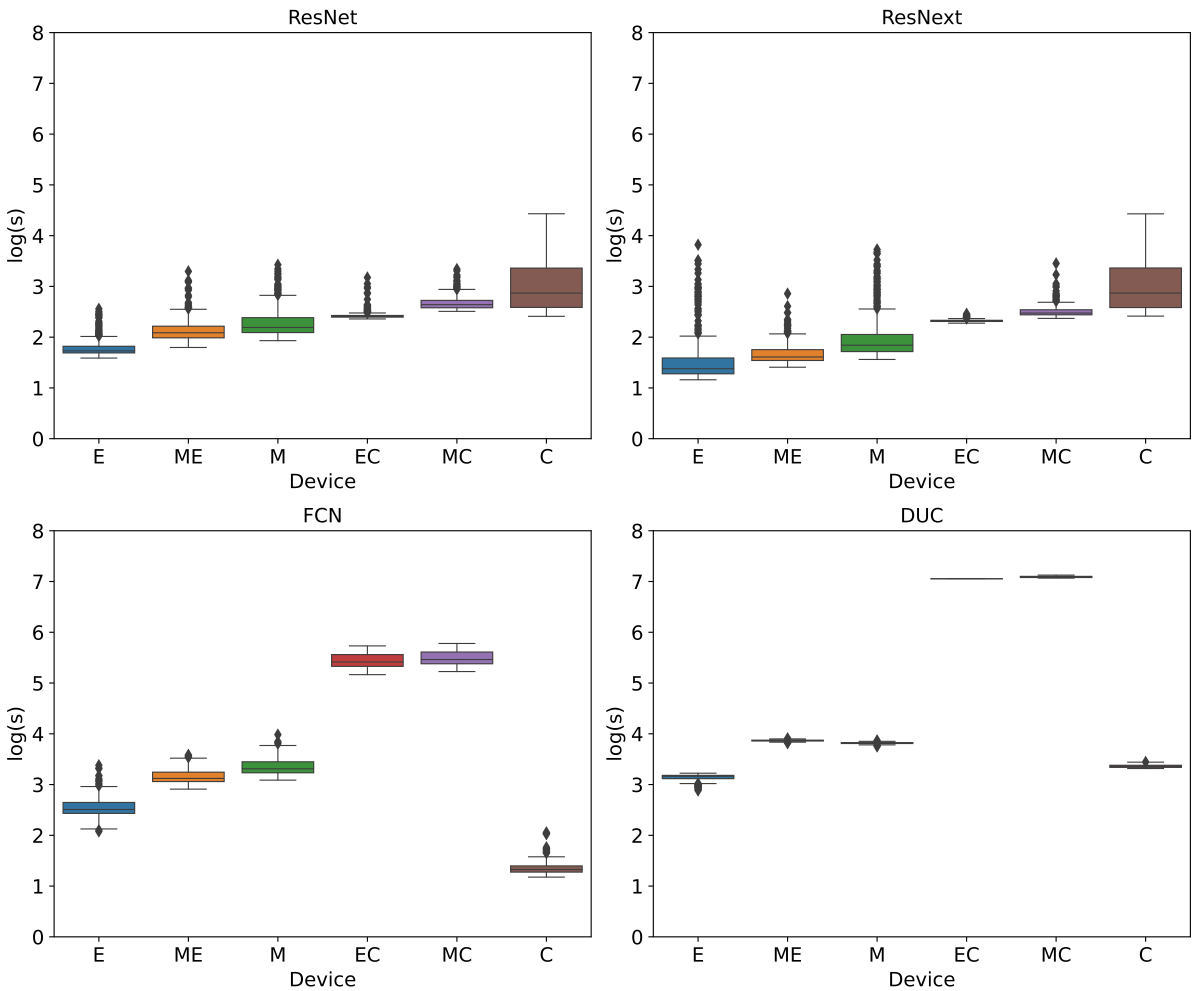


Fig. 6: Box plots of the inference latency measure for Multi-tier (Mobile-Edge [ME], Edge-Cloud [EC], Mobile-Cloud [MC]) Partitioned strategies and Single-tier (Mobile [M], Edge [E], Cloud [C]) monolithic deployment strategies

5.4 What is the impact of the Partitioned operator in terms of inference latency and accuracy across the considered tiers? (RQ4)

5.4.1 Data Exploration

Among the three Multi-tier Partitioned strategies, the Mobile-Edge Partitioned strategy shows 9.37x and 9.86x lower average median inference latency compared to Edge-Cloud and Mobile-Cloud Partitioned strategies, respectively.

The Mobile-Edge Partitioned strategy outperforms the Edge-Cloud and Mobile-Cloud Partitioned strategies in terms of median inference latency as depicted in Figure 6 (blue, orange, and green box-plots). In particular, it achieves a significant speedup, ranging from 1.37x to 24.17x and 1.74x to 24.95x compared to the Edge-Cloud and Mobile-Cloud Partitioned strategies (Table 29), respectively, with large effect sizes (Table 18).

The primary reason for this behavior is the impact of the size of intermediate data during transmission across the Edge-Cloud network. In the Edge-Cloud and Mobile-Cloud Partitioned strategies, the intermediate results generated after the first half of inference need to be transmitted from the Edge tier to the Cloud tier for the second half of inference. This transmission incurs additional latency, particularly when the intermediate data size of subjects (i.e., 6.12Mb for ResNet/ResNext, 118.12 to 200Mb for FCN, 781.25Mb for DUC) is larger than the limited Cloud network bandwidth (1 Mbps). Conversely, the Mobile-Edge Partitioned strategy is less influenced by the size of the intermediate data due to the higher Mobile-Edge network bandwidth (200 Mbps). Hence, the impact of the intermediate data size on inference latency is a crucial factor contributing to the observed performance advantage of the Mobile-Edge Partitioned strategy.

On the other hand, the Edge-Cloud Partitioned strategy achieves a speedup ranging from 1.03x to 1.26x when compared to the Mobile-Cloud Partitioned strategy, as shown in Figure 6 (yellow and green box-plots), along with small or large effect sizes (Table 18). These differences can be attributed to the faster computational capabilities of the Edge-Cloud tier compared to the Mobile-Cloud tier.

Table 18: RQ4 Results of the Cliff's Delta effect size and Conover test p-value between Mobile-Edge (ME), Edge-Cloud (EC), and Mobile-Cloud (MC) Partitioned strategies

P_x \ P_t	ME		EC		MC	
	p	δ	p	δ	p	δ
ME	-	-	$1.8e^{-117}$	L (0.77)	0.0	L (0.94)
EC	$8.1e^{-192}$	L (0.98)	-	-	$1.3e^{-148}$	L (0.96)
MC	0.0	L (0.99)	$1.0e^{-90}$	L (0.99)	-	-

P_d \ P_f	ME		EC		MC	
	p	δ	p	δ	p	δ
ME	-	-	0.0	L (1.0)	0.0	L (1.0)
EC	0.0	L (1.0)	-	-	$4.4e^{-20}$	S (0.27)
MC	0.0	L (1.0)	0.0	L (1.0)	-	-

[1] P_t, P_x, P_f, P_d denotes Partitioned versions of ResNet, ResNext, FCN, and DUC models, respectively.
[2] L, M, S, and N symbols denote Large, Medium, Small, and Negligible effect sizes, respectively.
[3] Empty cells indicate non-significant comparisons based on the Conover test.

The Edge Identity deployment strategy shows a 1.63x lower average median latency than the Mobile-Edge Partitioned strategy, which shows a 1.13x lower average median inference latency compared to the Mobile Identity deployment strategy.

The Edge Identity deployment strategy outperforms the Mobile-Edge Partitioned strategy, exhibiting a speedup ranging from 1.26x to 2.02x, as shown in Figure 6 (pink and dark blue boxplots), along with large effect sizes (Table 19). One possible explanation for this behavior is tier heterogeneity. In the Mobile-Edge deployment scenario, where distributed inference of partitioned models takes place, the tiers involved possess varying computation capabilities. Consequently, the processing speeds may vary, with slower tiers (such as Mobile) impacting the overall inference latency. This suggests that the use of a Partitioned model may not be necessary for scenarios where a monolithic deployment tier, such as Edge, is sufficiently capable of handling the computational load of the entire model.

In turn, the Mobile-Edge Partitioned strategy for all subjects (except DUC) achieves a speedup ranging from 1.11x to 1.26x compared to the Mobile Identity deployment strategy, as shown in Figure 6 (dark blue and red box plots), along with medium to large effect sizes (Table 19). Given that the Mobile tier in our study has the lowest computational resources compared to the more powerful Edge tier, offloading half of the computational load to the Edge tier alleviates the burden on the Mobile tier, resulting in reduced inference latency. These findings suggest that deploying partitioned models across resource-constrained tiers (i.e., Mobile and Edge) is more effective than deploying the entire model solely on the Mobile tier, due to the distribution of computational load during inference. For the DUC subject, the Mobile-Edge Partitioned strategy shows a 1.05x lower median inference latency than the Mobile Identity deployment strategy, with a large effect size. This is possible because of the large intermediate data size (781.25Mb) of the DUC model during distributed inference, which led to transmission overhead, even across the high Mobile-Edge network bandwidth of 200 Mbps, leading to a slower latency than the Mobile tier.

For ResNet/ResNext subjects, the Mobile-Edge, Edge-Cloud, and Mobile-Cloud strategies show 2.85x, 1.65, and 1.36 lower average median latency, respectively compared to the Cloud Identity deployment strategy. In contrast, in the case of FCN/DUC subjects, they show a 3.84x, 49.98x, and 52.14x higher average median inference latency.

For ResNet/ResNext subjects, the Mobile-Edge, Edge-Cloud, and Mobile-Cloud Partitioned strategies achieve a speedup of 2.18x to 3.52x, 1.58x to 1.73x, and 1.25x to 1.48x, respectively, compared to the Cloud Identity deployment strategy, as illustrated in Figure 6, along with medium to large effect sizes (Table 19). The lower intermediate data size (6.12 Mb) in comparison to the input data size (8 to 60 Mb) of these two subjects speeds up their transmission across the Mobile, Edge, and Cloud tiers for the 3 Multi-tier Partitioned strategies. It suggests that Partitioned strategies can be a better alternative than Cloud deployment for subjects having intermediate data sizes lower than the input data.

For FCN/DUC subjects, the Mobile-Edge, Edge-Cloud, and Mobile-Cloud Partitioned strategies exhibit 1.67x to 6.01x, 40.38x to 59.59x, and 41.68x to 62.60x higher median inference latency, respectively, than the Cloud Identity deployment strategy, with large effect sizes (Table 19). For DUC, the intermediate data size (781.25 Mb) is much higher in comparison to the input data size (19-22 Mb), which led to its transmission overhead across both the Mobile-Edge (200 Mbps) and Edge-Cloud (1 Mbps) networks for the 3 Multi-tier Partitioned strategies. Conversely, for FCN, the intermediate data size varies between 118.12 Mb to 200 Mb, which is still much higher than their input data size (2-5 Mb), leading to transmission overhead. This is especially the case for the Edge-Cloud and Mobile-Cloud Partitioned strategies, which require transmission across the constrained Edge-Cloud network (1 Mbps). For the same subject, the slower latency of Mobile-Edge Partitioned than of Cloud deployment is majorly due to the computational advantage of the Cloud compared to the Mobile-Edge tier. This suggests that Cloud deployment can be a better alternative than Partitioned strategies for subjects having input data sizes smaller than the intermediate data. Moreover, for the FCN subject, which has the smallest input data sizes among the four subjects, its Cloud Identity deployment strategy also shows faster latency than its Mobile/Edge Identity deployment strategies, as explained in RQ1 findings (Section 5.1).

The Edge Early Exit/Quantized deployment strategy shows 1.96x/2.41x lower average median latency than the Mobile-Edge Partitioned strategy at a medium/small accuracy loss.

The Edge Early Exit/Quantized deployment strategy shows lower median inference latency than the Mobile-Edge Partitioned strategy, ranging from 1.67x to 2.35x/ 1.87x to 2.44 across the four subjects, as shown in Figure 5 (blue, orange, green box-plots), along with large effect sizes (Table 12). In terms of accuracy, the Early Exit and Quantized operators show medium (2.53% to 11.56%) and small (0.05% to 0.38%) accuracy loss relative to Identity (or Partitioned) models as stated in RQ2 and RQ3 findings. This indicates that the Quantized and Early Exit operators at the Edge tier are a better alternative than the Partitioned operator at the Mobile-Edge tier in scenarios where sacrificing a small to medium level of accuracy may be acceptable to achieve faster latency. Table 19 shows a summary of results for RQ4.

5.4.2 Normality Test

To assess the applicability of parametric statistical tests, the Shapiro-Wilk test was conducted for each subject-model pair across ME, EC, and MC strategies. In all cases, p-values were less than 0.05, indicating deviations from normality (Table 29). This conclusion was further supported by the Q-Q plots (Table 30), which displayed noticeable divergence from the expected linear pattern. Therefore, non-parametric tests were employed in subsequent analyses.

5.4.3 Hypothesis Testing

All partitioning strategies (ME, EC, MC) differ statistically and practically in how they affect model performance in terms of pairwise comparisons using the Conover test, as shown in Table 18.

Table 19: RQ4 results of the Cliff's Delta effect size and Conover test p-value between Multi-tier (Mobile-Edge [ME], Edge-Cloud [EC], Mobile-Cloud [MC]) Partitioned Strategies and Single-tier (Mobile [M], Edge [E], Cloud [C]) Monolithic Strategies.

P_t \ I_t	M		E		C	
	p	δ	p	δ	p	δ
ME	$2.0e^{-36}$	-M (0.37)	$9.4e^{-112}$	L (0.83)	0.0	-L (0.95)
EC	$3.5e^{-29}$	L (0.54)	0.0	L (0.96)	$2.0e^{-221}$	-L (0.95)
MC	$1.8e^{-267}$	L (0.73)	0.0	L (1.0)	$1.1e^{-12}$	-M (0.33)

P_x \ I_x	M		E		C	
	p	δ	p	δ	p	δ
ME	$9.0e^{-77}$	-L (0.63)	$9.1e^{-12}$	L (0.53)	0.0	-L (1.0)
EC	$2.9e^{-36}$	L (0.68)	$7.9e^{-266}$	L (0.82)	$4.2e^{-241}$	-L (1.0)
MC	$8.3e^{-211}$	L (0.74)	0.0	L (0.84)	$1.7e^{-52}$	-L (0.76)

P_f \ I_f	M		E		C	
	p	δ	p	δ	p	δ
ME	$2.0e^{-103}$	-L (0.65)	$1.0e^{-269}$	L (0.98)	0.0	L (1.0)
EC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)
MC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)

P_d \ I_d	M		E		C	
	p	δ	p	δ	p	δ
ME	0.0	L (0.98)	0.0	L (1.0)	0.0	L (1.0)
EC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)
MC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)

1 P_t, P_x, P_f, P_d denote Partitioned versions of ResNet, ResNext, FCN, and DUC respectively.
2 I_t, I_x, I_f, I_d denote Identity versions of ResNet, ResNext, FCN, and DUC respectively.
3 A Positive sign for the i^{th} cell means latency of column[i] < row[i], and Negative means column[i] > row[i].
4 L, M, S, and N symbols mean Large, Medium, Small, and Negligible effect sizes, respectively.
5 Empty cells (not shown) would indicate non-significant results based on the Conover test ($p > 0.05$).

Summary of Research Question 4

The Edge Identity/Early Exit/Quantized deployment strategy shows faster latency (1.63x/1.96x/2.41x) at no/medium/small accuracy loss than the ME Partitioned strategy, which exhibits faster latency (1.13x) compared to the Mobile Identity deployment strategy (1.13x), EC/MC Partitioned strategy (9.37x/9.86x), and Cloud Identity deployment strategy (2.85x for ResNet/ResNext) in deployment scenarios where factors like input/intermediate data size and computational/network resources play a crucial role.
In scenarios where the subjects have smaller input data sizes (i.e., FCN) such that their transmission across the bandwidth-constrained Cloud tier is not a major concern, their monolithic Cloud Identity deployment is much more effective than their Multi-tier Partitioned strategies and Edge/Mobile Identity deployment strategies.

Model	Best Partitioning Tier	Latency Gain	Notes
ResNet	Mobile-Edge	1.13x faster than Mobile	9.37x faster than EC, 9.86x faster than MC; Edge Identity is still faster (1.63x)
ResNext	Mobile-Edge	1.13x faster than Mobile	Similar trend as ResNet; large intermediate data size limits EC/MC strategies
FCN	Cloud	Cloud outperforms ME	Small input size makes Cloud preferable over any partitioned strategy
DUC	Edge	Edge Identity faster (1.96x)	Intermediate data too large (781MB) to benefit from partitioning

Table 20: Summary of results for Research Question 4

5.5 What is the impact of Hybrid operators in terms of inference latency and accuracy within and across the considered tiers? (RQ5)

5.5.1 Data Exploration

5.5.2 Quantitative Analysis of Quantized Early Exit operator on monolithic deployment tiers

In Mobile, for two subjects (FCN/DUC), the Quantized Early Exit models show 1.45x, 1.13x, and 1.35x lower average median inference latency than the Identity, Quantized, and Early Exit models, respectively. In Edge, for all subjects, the Quantized Early Exit models show 1.75x, 1.17x, and 1.45 lower average median inference latency than the Identity, Quantized, and Early Exit models, respectively. In Cloud, the Quantized Early Exit models show no significant difference from the Identity, Quantized, and Early Exit models for all subjects (except DUC).

In the Mobile tier, for FCN and DUC, the Quantized Early Exit models show 1.29x to 1.62x and 1.16x to 1.54x lower median inference latency than the Identity and Early Exit models, respectively (Figure 5) along with large effect sizes (Table 11, Figure 29, Figure 26, Figure 28). Conversely, for the ResNet/ResNext subject, the Quantized Early Exit models show 1.07x lower/1.05x higher (small effect sizes) and 1.11x (negligible effect size)/1.33x (large effect size) higher median inference latency than the Identity and Early Exit models, respectively. Even though the Quantized Early Exit models have 4.61x to 5.04x and 3.90x to 3.98x lower size than the Identity and Early Exit models respectively, still they show slower latency for ResNet, ResNext, or both during Mobile deployment, which indicates that these two subjects' Quantization Operations are costly in lower Memory/CPU environments, as mentioned in RQ2 findings (Section 5.2). Moreover, the Quantized Early Exit models show 1.09x to 1.28x lower median inference latency than the Quantized models across the four subjects, with small to large effect sizes. This is possibly due to the addition of early exiting, which reduces computations during inference in comparison to a Quantized variant without any early exit similar to RQ3 findings (Section 5.3).

During Edge deployment, across the four subjects, the Quantized Early Exit models exhibit 1.36x to 1.97x, 1.13x to 1.20x, 1.17x to 1.82x lower median inference latency than the Identity, Quantized, and Early Exit models, respectively, as depicted in Figure 5, with large effect sizes (Table 12). This suggests that the Quantized Early Exit models exhibit greater robustness on the Edge tier in comparison to the Mobile tier. The higher computational resource on the Edge tier is likely a contributing factor to this outcome.

For the DUC subject, the Quantized Early Exit models show slower latency than the Identity and Early Exit models with large effect sizes (Table 13). We believe that this is due to the lowest percentage of graphical nodes processed with the CUDA Execution Provider for DUC during Cloud deployment, similar to the reasoning explained briefly for the RQ2 results (see first finding in Section 5.2). For the same subject, the Quantized Early Exit model shows faster latency than the Quantized model, but with a small effect size, possibly due to the minor influence of Early Exiting.

The Quantized Early Exit models show a medium drop in accuracy relative to Identity, Quantized, and Early Exit models.

In Table 9, an accuracy drop of 2.62% to 12.41%, 2.56% to 12.02%, 0.09% to 1.66% is observed when comparing the performance of the Quantized Early Exit models with the Identity, Quantized, and Early Exit models, respectively. These accuracy differences are statistically significant (Wilcoxon Test: p-value = 7.6e^{-6}, $\alpha = 0.0125$), with medium effect sizes. For applications prioritizing real-time inference in the

Edge tier and willing to accept a medium decrease in accuracy compared to the Identity, Quantized, and Early Exit models, the utilization of a Hybrid (Quantized Early Exit) model may be a suitable choice.

When comparing Quantized Early Exit across the three monolithic deployment tiers, the Quantized Early Exit models during Edge deployment show 2.63x and 3.89x lower average median inference latency than their deployment in Mobile and Cloud tiers, respectively.

The Quantized Early Exit models in the Edge tier demonstrate 2.03x to 2.90x lower median inference latency compared to the Mobile tier as shown in Figure 5 (red box plots), with large effect sizes (Table 21) due to Edge's higher computational resources. Among the Edge and Cloud tiers, the Quantized Early Exit model's median inference latency under-performs by 1.77x to 7.69x during Cloud deployment for three subjects (i.e., ResNet, ResNext, DUC), along with a large effect size (Table 21). This is due to the higher impact of large input data sizes of these subjects on the transmission across the restricted Edge-Cloud network during Cloud deployment. Conversely, for the FCN subject, the Quantized Early Exit model exhibits 1.93x lower median inference latency in the Cloud tier compared to the Edge tier, with a large effect size (Table 21). Similarly, this is due to FCN's small input data sizes, leading to a lower impact on data transmission across the constrained Edge-Cloud network during Cloud deployment. The reasoning behind these findings is similar to and explained briefly in the RQ1 findings (Section 5.1).

Table 21: RQ5 Results of the Cliff's Delta effect size and Conover test p-value between Mobile (M), Edge (E), and Cloud (C) deployment for Quantized Early Exit models

QE_{x} \ QE_{t}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	$2.5e^{-304}$	L (0.99)	$9.4e^{-260}$	-L (0.91)
E	0.0	L (1.0)	-	-	0.0	-L (1.0)
C	0.0	-L (1.0)	0.0	-L (1.0)	-	-

QE_{d} \ QE_{f}	M		E		C	
	p	δ	p	δ	p	δ
M	-	-	0.0	L (1.0)	0.0	L (1.0)
E	0.0	L (1.0)	-	-	0.0	L (0.97)
C	$8.7e^{-316}$	L (0.97)	0.0	-L (1.0)	-	-

[1] QE_{t}, QE_{x}, QE_{f}, and QE_{d} denote Quantized Early Exit versions of ResNet, ResNext, FCN, and DUC, respectively.
[2] L, M, S, and N symbols denote Large, Medium, Small, and Negligible effect sizes, respectively.
[3] Empty cells indicate non-significant comparisons based on the Conover test.

5.5.3 Quantitative Analysis of Quantized Early Exit Partitioned Strategy and its Comparison with monolithic deployment Strategies

Across the three Multi-tier Quantized Early Exit Partitioned strategies, the Mobile-Edge Quantized Early Exit Partitioned strategy shows an 8.51x and 9.04x lower average median inference latency than Edge-Cloud and Mobile-Cloud Quantized Early Exit Partitioned strategies, respectively.

The Mobile-Edge Quantized Early Exit Partitioned strategy accounts for 1.73x to 15.65x and 2.21x to 16.20x lower median inference latency than the Edge-Cloud and Mobile-Cloud Quantized Early Exit Partitioned strategies, respectively (Figure 7), with large effect size (Table 22). On the other hand, the Edge-Cloud Quantized Early Exit Partitioned strategy shows 1.04x to 1.28x lower median inference latency than the Mobile-Cloud Quantized Early Exit Partitioned strategy as shown in Figure 7, along with small or large effect sizes (Table 22). This is due to similar reasons as explained in the RQ4 results (see first finding in Section 5.4).

The Edge Quantized Early Exit deployment strategy shows 2.03x lower average median inference latency than the Mobile-Edge Quantized Early Exit Partitioned strategy, which shows 1.29x lower average median inference latency than the Mobile Quantized Early Exit deployment strategy.

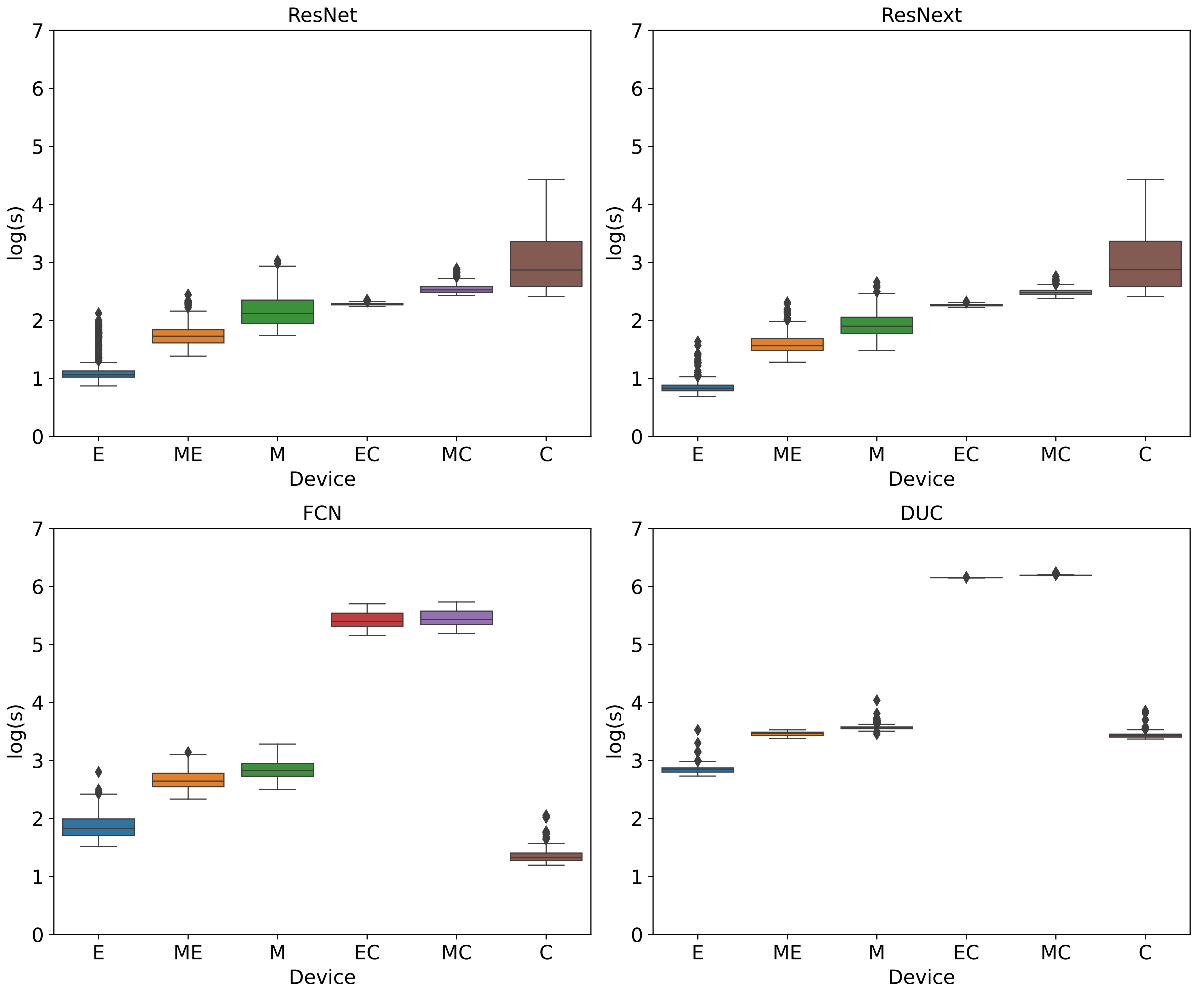


Fig. 7: Box plots of the inference latency measure for Multi-tier Quantized Early Exit Partitioned strategies and monolithic Quantized Early Exit deployment strategies involving 6 deployment tiers (i.e., Mobile-Edge [ME], Edge-Cloud [EC], Mobile-Cloud [MC], Mobile [M], Edge [E], Cloud [C])

The Edge Quantized Early Exit strategy outperforms the Mobile-Edge Quantized Early Exit Partitioned strategy with large effect sizes (Table 23), exhibiting a speedup ranging from 1.85x to 1.94x, as shown in Figure 7. In turn, the Mobile-Edge Quantized Early Exit Partitioned strategy shows 1.09x to 1.47x lower median inference latency than the Mobile Quantized Early Exit strategy (large effect sizes) as shown in Figure 7 and Table 23. In general, this suggests that distributed model deployment across Mobile and Edge tiers can be an optimal choice compared to monolithic model deployment in a resource-constrained environment (such as mobile) when faster inference is a concern at no accuracy loss. This is due to similar reasons provided in RQ4 results (see second finding in Section 5.4).

For ResNet/ResNext, the Mobile-Edge, Edge-Cloud, and Mobile-Cloud Quantized Early Exit Partitioned strategies show 3.41x, 1.82x, and 1.44x lower average median latency, respectively than the Cloud Quantized Early Exit deployment strategy. In contrast, in the case of FCN/DUC, they show 2.39x, 36.92x, and 38.28x higher average median inference latency.

For ResNet and ResNext, the Mobile-Edge, Edge-Cloud, and Mobile-Cloud Quantized Early Exit Partitioned strategies demonstrate a lower median inference latency of 3.12x to 3.70x, 1.80x to 1.84x, and 1.40x to 1.48x, respectively than the Cloud Quantized Early Exit strategy as shown in Figure 7, along with large effect sizes (Table 23). Conversely, for FCN and DUC, the same strategies exhibit a 1.04x to 3.74x, 15.28x to 58.57x, and 15.91x to 60.65x higher median inference latency, along with medium to large effect sizes (Table 23). A possible explanation for these results is similar to the RQ4 results (see

Table 22: RQ5 Results of the Cliff's Delta effect size and Conover test p-value between Mobile-Edge (ME), Edge-Cloud (EC), and Mobile-Cloud (MC) for Quantized Early Exit Partitioned Strategies

QEP_x \ QEP_t	ME		EC		MC	
	p	δ	p	δ	p	δ
ME	-	-	$3.6e^{-305}$	-L (-0.97)	0.0	-L (1.0)
EC	0.0	-L (0.99)	-	-	$7.0e^{-241}$	-L (1.0)
MC	0.0	-L (1.0)	$1e^{-323}$	-L (1.0)	-	-

QEP_d \ QEP_f	ME		EC		MC	
	p	δ	p	δ	p	δ
ME	-	-	0.0	-L (1.0)	0.0	-L (1.0)
EC	0.0	-L (1.0)	-	-	$3.8e^{-16}$	-S (0.25)
MC	0.0	-L (1.0)	$1.5e^{-268}$	-L (1.0)	-	-

[1] QEP_t, QEP_x, QEP_f, and QEP_d denote Quantized Early Exit Partitioned versions of ResNet, ResNext, FCN, and DUC, respectively.
[2] L, M, S, and N symbols denote Large, Medium, Small, and Negligible effect sizes, respectively.
[3] All p-values shown are statistically significant ($p < 0.05$); empty cells denote non-significant comparisons.

third finding in Section 5.4). Moreover, for the FCN subject, which has the lowest input data sizes among the four subjects, its Cloud Quantized Early Exit deployment strategy also shows faster latency than its Mobile/Edge Quantized Early Exit deployment strategies, as explained previously in RQ5 findings (Section 5.5.1).

The Edge Identity/Quantized/Early Exit deployment strategy shows 1.17x/1.72x/1.42x lower average median latency at medium accuracy gain when compared with Mobile-Edge Quantized Early Exit Partitioned strategy, which shows 1.39x lower average median latency than the Mobile-Edge Partitioned strategy at medium accuracy loss.

The Edge Identity/Quantized/Early Exit deployment strategies show 1x to 1.35x/ 1.61x to 1.88x/ 1.24x to 1.59x lower median inference latency than the Mobile-Edge Quantized Early Exit Partitioned strategy across the four subjects, as shown in Figure 5 (blue, orange, green, red box-plots), along with medium to large effect sizes (Table 12). In terms of accuracy, the Identity, Quantized, and Early Exit models show medium accuracy gain (i.e., 2.62% to 12.41%, 2.56% to 12.02%, 0.09% to 1.66%) relative to Quantized Early Exit (or Quantized Early Exit Partitioned) models as stated previously in the second finding of RQ5 (Section 5.5.2). This indicates that the Non-Hybrid operators (Identity, Quantized, and Early Exit) at the Edge tier are a better alternative than the Hybrid Quantized Early Exit Partitioned operator at the Mobile-Edge tier in scenarios where maximizing both latency and accuracy are of utmost importance. Moreover, the Mobile-Edge Quantized Early Exit Partitioned strategy shows 1.04x to 1.60x lower median inference latency with large effect sizes (Table 12) than the Mobile-Edge Partitioned strategy but at a cost of medium accuracy loss (2.62% to 12.41%). Table 23 shows a summary of results for RQ5.

5.5.4 Normality Test

The Shapiro-Wilk test and QQ plots (Table 30, Figure 31, Figure 32) suggest that latency data for most deployment tiers and model variants deviate from normality. P-values are < 0.05 across all cases. This non-normality validates the use of non-parametric tests (Conover and Cliff's Delta).

5.5.5 Hypothesis Testing

According to the Conover test, the null hypothesis that there is no significant difference between Quantized Early Exit and Identity, Quantized Early Exit and Quantized, Quantized Early Exit and Early Exit models in the Cloud cannot be rejected for three subjects (i.e., ResNet, ResNext, FCN), indicating their similar or equivalent latency performance in the Cloud. Due to the powerful computing resources of the Cloud, the impact of the Quantized Early Exit models is not significant compared to the Identity,

Table 23: RQ5 Results of the Cliff's Delta effect size and Conover test p-value between Multi-tier (Mobile-Edge [ME], Edge-Cloud [EC], Mobile-Cloud [MC]) Quantized Early Exit Partitioned Strategies and Single-tier (Mobile [M], Edge [E], Cloud [C]) Quantized Early Exit strategies

QEP_t \ QE_t	M		E		C	
	p	δ	p	δ	p	δ
ME	$7.9e^{-228}$	-L (0.85)	$3.7e^{-130}$	L (0.90)	0.0	-L (1.0)
EC	$5.9e^{-12}$	M (0.35)	0.0	L (1.0)	0.0	-L (1.0)
MC	$5.6e^{-319}$	L (0.78)	0.0	L (1.0)	$6.3e^{-73}$	-L (0.65)

QEP_x \ QE_x	M		E		C	
	p	δ	p	δ	p	δ
ME	$1.0e^{-213}$	-L (0.82)	$5.4e^{-287}$	L (0.99)	0.0	-L (1.0)
EC	$7.7e^{-218}$	L (0.85)	0.0	L (1.0)	0.0	-L (1.0)
MC	0.0	L (0.98)	0.0	L (1.0)	$1.4e^{-172}$	-L (0.80)

QEP_f \ QE_f	M		E		C	
	p	δ	p	δ	p	δ
ME	$1.1e^{-76}$	-L (0.57)	$2.3e^{-290}$	L (1.0)	0.0	L (1.0)
EC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)
MC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)

QEP_d \ QE_d	M		E		C	
	p	δ	p	δ	p	δ
ME	0.0	-L (0.99)	0.0	L (1.0)	$4.2e^{-55}$	M (0.42)
EC	$2.1e^{-277}$	L (1.0)	0.0	L (1.0)	0.0	L (1.0)
MC	0.0	L (1.0)	0.0	L (1.0)	0.0	L (1.0)

[1] QEP_t, QEP_x, QEP_f, QEP_d denote Quantized Early Exit Partitioned versions of ResNet, ResNext, FCN, and DUC, respectively.
[2] QE_t, QE_x, QE_f, QE_d denote single-tier Quantized Early Exit versions of the same models.
[3] L, M, S, and N denote Large, Medium, Small, and Negligible Cliff's Delta effect sizes.
[4] All p-values shown are significant ($p < 0.05$).

Quantized, and Early Exit models. However, for the DUC subject, the null hypothesis was rejected with p-values of $8.9e^{-221}$, $6.8e^{-10}$ and $1.5e^{-196}$ (α =0.05) for Quantized Early Exit vs Identity, Quantized Early Exit vs Quantized, and Quantized Early Exit vs Early Exit comparison, respectively, indicating significant inference latency difference between them.

Summary of Research Question 5

The Quantized Early Exit operator shows 1.75x/1.17x/1.45 faster latency than the Identity/Quantized/Early Exit operator in the Edge tier at medium accuracy loss (up to 12.41%/12.02%/1.66%). The Edge deployment of Hybrid (Quantized Early Exit) and Non-Hybrid (Identity/Quantized/Early Exit) operators show 2.03x and 1.17x/1.72x/1.42x faster latency at no accuracy loss and medium (up to 12.41%/12.02%/1.66%) accuracy gain, respectively than the ME Quantized Early Exit Partitioned strategy, which shows faster latency than the ME Partitioned strategy (1.39x at medium accuracy loss), EC/MC Quantized Early Exit Partitioned (8.51x/9.04x), Mobile Quantized Early Exit (1.29x), & Cloud Quantized Early Exit (3.41x for ResNet/ResNext) strategies for scenarios influenced by input/intermediate data size of the subjects and computational/bandwidth resources of the Mobile, Edge, and Cloud tiers.
In scenarios where the subjects have smaller input data sizes (i.e., FCN) such that their transmission across the bandwidth-constrained Cloud tier is not a major concern, their monolithic Cloud Quantized Early Exit deployment is much more effective than their Multi-tier Quantized Early Exit Partitioned strategies and Edge/Mobile Quantized Early Exit deployment strategies.

Table 24: Summary of results for Research Question 5

Model	Best Tier	Latency Gain	Accuracy Drop	Notes
ResNet	Edge	1.75x vs Identity; 2.03x vs ME-QEP	12.41%	Best performance on Edge. ME-QEP faster than ME-Part (1.39x)
ResNext	Edge	1.17x–1.72x vs Non-Hybrids; 2.03x vs ME-QEP	12.02%	Similar trends to ResNet; large effect sizes in latency gains
FCN	Cloud	1.93x faster than Edge for QE	2.62%	Cloud deployment preferable due to small input size
DUC	Edge	1.45x faster vs Early Exit; 2.03x vs ME-QEP	1.66%	High input size favors Edge deployment; ME-QEP adds moderate accuracy loss

5.6 What is the impact of network bandwidth variations on the deployment strategies in terms of inference latency? (RQ6)

5.6.1 Data Exploration

Deploying smaller input data-sized models (FCN), is well-suited for Cloud tier with lower bandwidth (≤10 Mbps). Larger input data-sized models (ResNe(x)t and DUC), perform better in Cloud deployments having moderate to high bandwidth (≥50 Mbps).

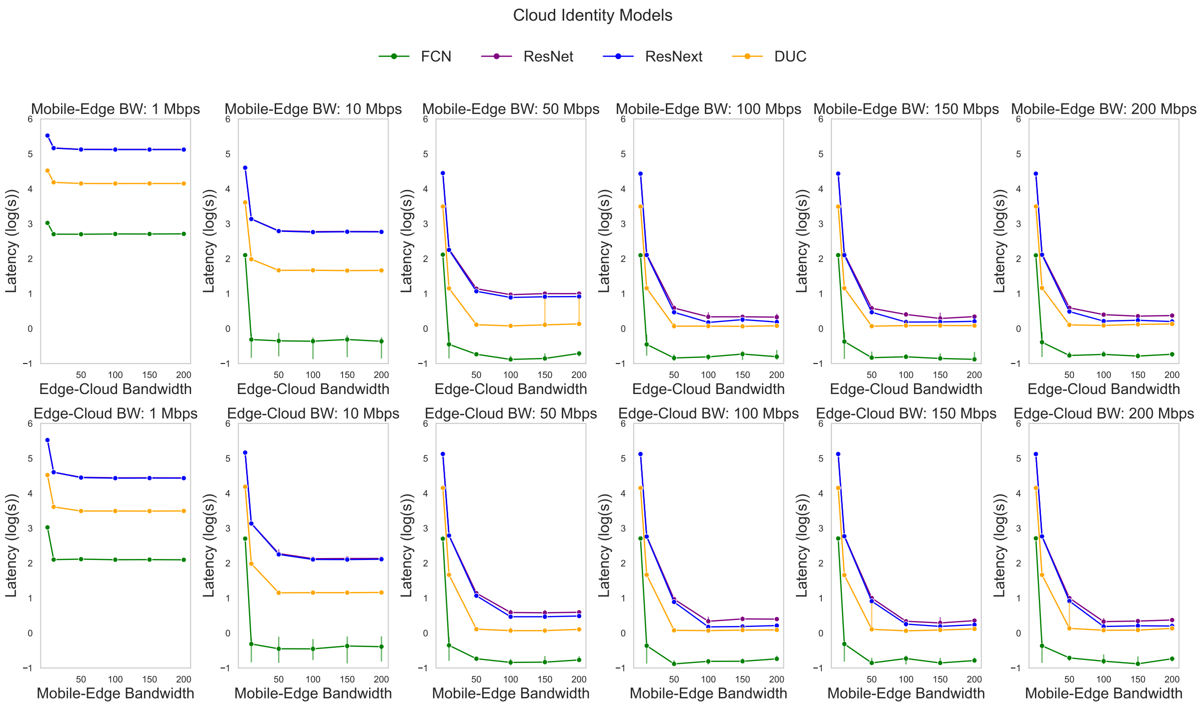


Fig. 8: Line graphs of the measures collected for inference latency vs network bandwidth of Cloud Identity models.

Figure 8 shows the line graphs of the measures collected for the latency of Identity models in the Cloud across the four subjects (i.e., FCN, ResNet, ResNext, DUC) for various network bandwidths. For the other models (Quantized, Early Exit, and Quantized Early Exit), similar line graphs were observed and therefore added to the replication package [8]. As both Mobile-Edge and Edge-Cloud bandwidth increases, latency tends to drop until either 50 or 100 Mbps, then stays constant. The latency performance gap between the models narrows as the bandwidth increases, indicating the models perform more similarly at higher bandwidth levels (≥100 Mbps). For all cases, a steep latency drop is observed initially when bandwidth increases from 1 to 50 Mbps. Beyond 100 Mbps, the latency changes become steady.

[8] https://github.com/SAILResearch/wip-24-jaskirat-black-box-edge-operators

FCN (green line) consistently shows the lowest latency across all scenarios in the Cloud, indicating its efficiency in handling workloads under varying bandwidth conditions when comparing the different plots in a row. This is because the FCN's input data size (5 Mb) is the smallest one across all the datasets. This suggests that deploying models with small input sizes like FCN is ideal for Cloud deployment even in bandwidth-constrained settings ($\leq$10 Mbps), which complements the RQ1-5 findings for FCN-based Cloud operators. ResNext and ResNet (blue and purple lines) models tend to have the highest latency among the models, especially in low network bandwidth conditions (i.e., $\leq$10 Mbps), which again aligns with the RQ1-5 findings. As the input data size of ResNe(x)t (i.e., 60 Mb) is much higher than the network-constrained bandwidth (i.e.,$\leq$10 Mbps), its latency is higher compared to other models at $\leq$10 Mbps. The DUC (orange line) demonstrates moderate latency compared to other models, such as FCN and ResNe(x)t in network-constrained scenarios (i.e., 10 Mbps). The DUC and ResNe(x)t models' latency performance reduces sharply as the bandwidth increases from 1 Mbps to 50 Mbps, eventually reaching a plateau at 50 or 100 Mbps, indicating a steady state. The input data size of DUC (22 Mb) and ResNe(x)t (60 Mb) exceeds the 10 Mbps bandwidth but falls below the 50 Mbps bandwidth (for DUC) and is slightly above the 50 Mbps bandwidth (for ResNe(x)t). As a result, latency decreases progressively for ResNe(x)t and DUC models as the bandwidth rises from 10 Mbps to 50 Mbps. This suggests that the ResNe(x)t and DUC models are suitable for Cloud deployment that operates in scenarios with at least moderate bandwidth availability ($\geq$50 Mbps) across Mobile-Edge and Edge-Cloud networks, but might not be ideal for constrained bandwidth settings ($\leq$10 Mbps), which complements the RQ1-5 findings for ResNe(x)t/DUC-based cloud models.

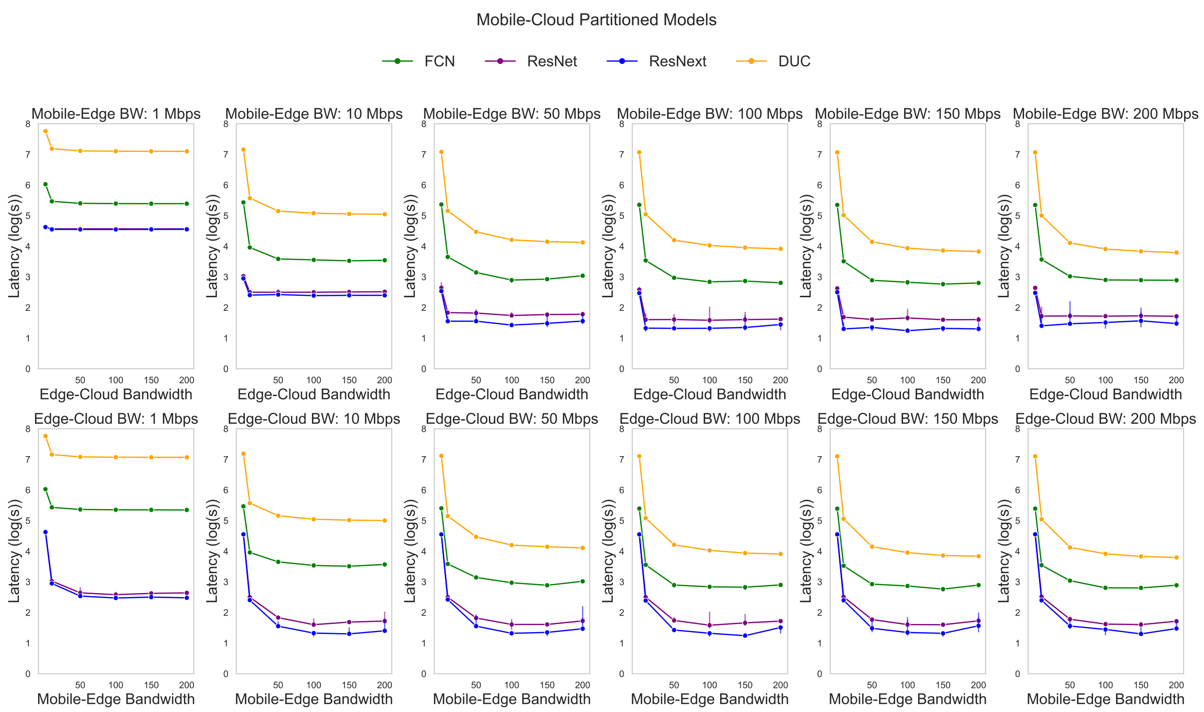


Fig. 9: Line graphs of the measures collected for inference latency vs network bandwidth of Mobile-Cloud Partitioned models.

For larger intermediate data-sized models (FCN, DUC), Partition-based strategies require $\geq$50 Mbps to achieve latency convergence. The Non-Partitioned models with large input data sizes (ResNe(x)t) are suitable for Mobile and Edge deployments at $\geq$50 Mbps.

The Figure 9, 10, 11,and 12 show the line graphs of the measures collected for latency vs bandwidth of Mobile-Cloud Partitioned, Mobile-Cloud Quantized Early Exit Partitioned, Edge-Cloud Partitioned, and Edge-Cloud Quantized Early Exit Partitioned models across the four subjects (i.e., FCN, ResNet, ResNext, DUC). For these Partitioned-based models, the top row shows the steady state of ResNe(x)t latency when Edge-Cloud bandwidth is varied and the Mobile-Edge bandwidth is kept fixed, possibly

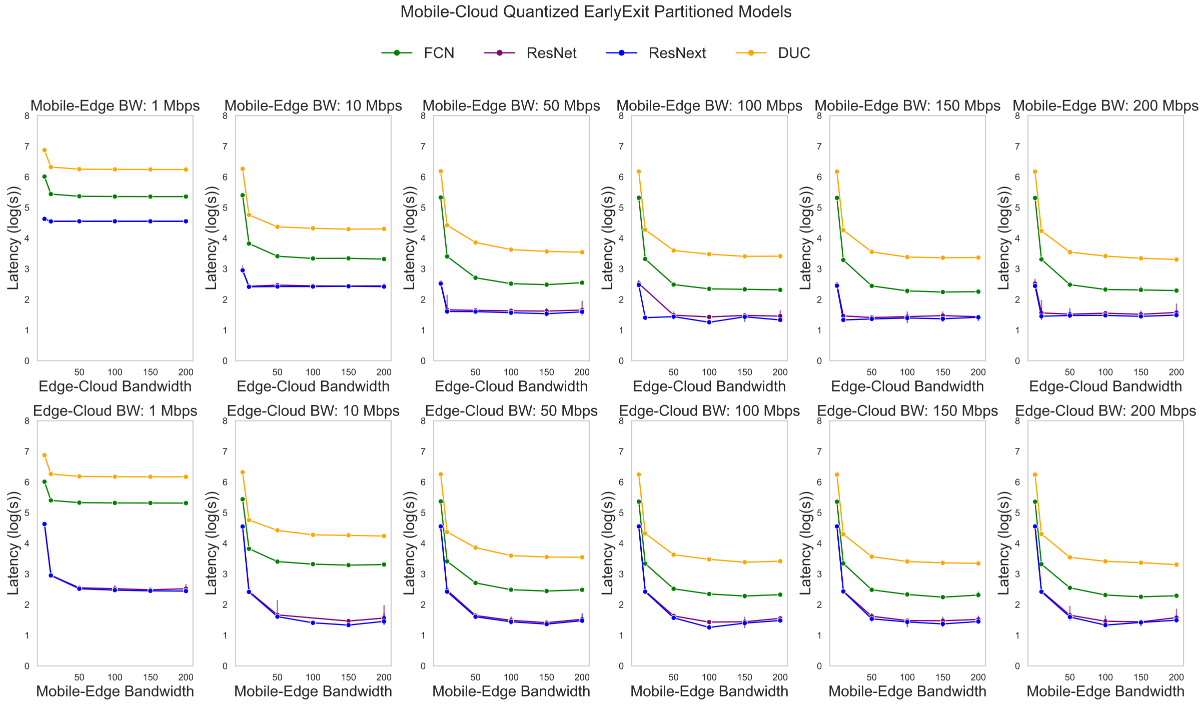


Fig. 10: Line graphs of the measures collected for inference latency vs network bandwidth of Mobile-Cloud Quantized Early Exit Partitioned models.

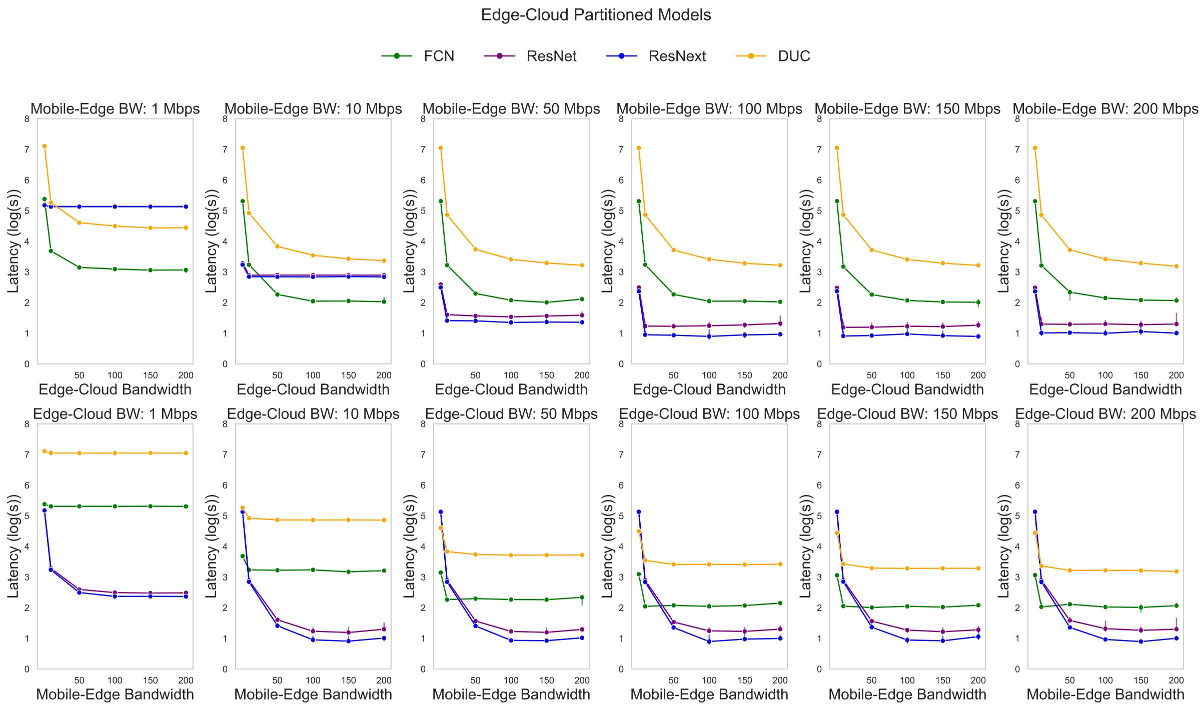


Fig. 11: Line graphs of the measures collected for inference latency vs network bandwidth of Edge-Cloud Partition models.

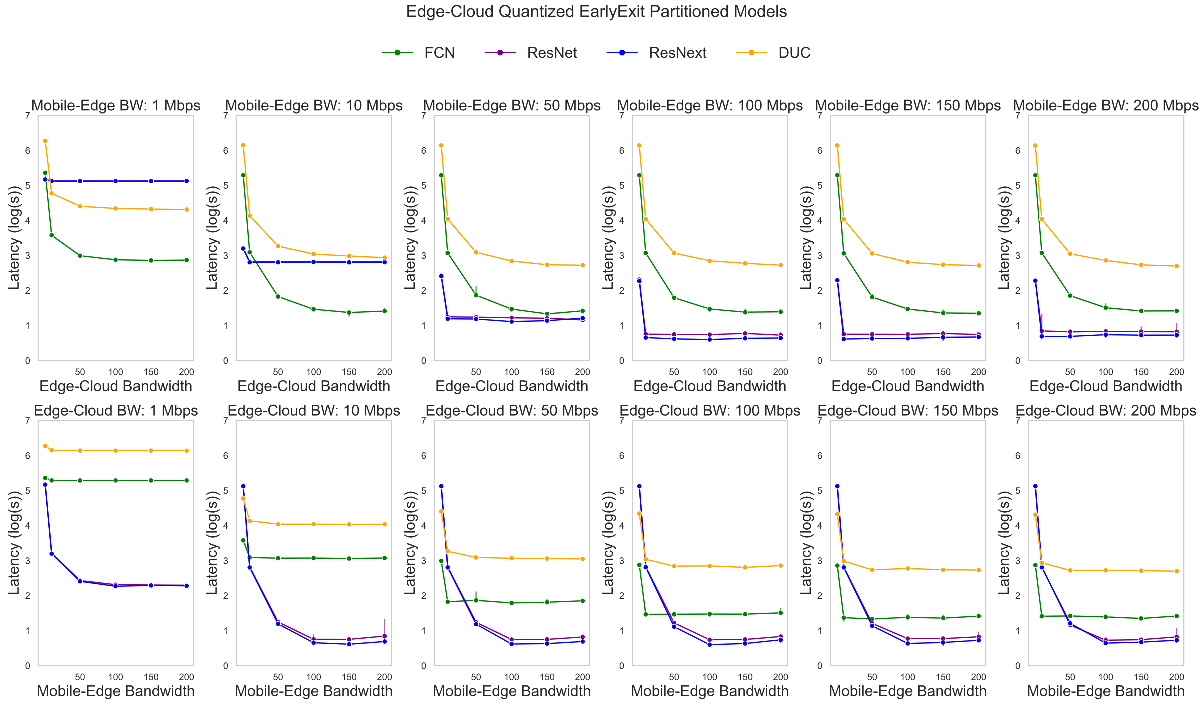


Fig. 12: Line graphs of the measures collected for inference latency vs network bandwidth of Edge-Cloud Quantized Early Exit Partitioned models.

due to the lower influence of intermediate data size (6.12 Mb) on Edge-Cloud bandwidths, similar to the observations in RQ4 and RQ5 findings (Section 5.4, 5.5.3) when analyzing Partitioning-based strategies. Moreover, the steady line graph (top row) drops to lower latency as the Mobile-Edge bandwidth increases from 1 to 50 Mbps. A sudden drop in latency is observed when increasing the Mobile-Edge bandwidth from 1 to 50 Mbps at fixed Edge-Cloud bandwidths (bottom row) because the influence of input data size (60 Mb) transmission decreases on the latency. For Mobile-Cloud Partitioned-based FCN models, the latency gradually decreases with the increase in either Edge-Cloud bandwidth (top row) or Mobile-Edge bandwidth (bottom row) and converges at 50 or 100 Mbps. However, for Edge-Cloud Partitioning-based FCN models, the latency plateaus at 50 or 100 Mbps of Edge-Cloud bandwidth at fixed Mobile-Edge bandwidths (top row), and the steady line graphs (bottom row) drops to lower latency as the Edge-Cloud bandwidth increases from 1 to 100 Mbps. The reason is the low impact of intermediate data size (135 Mb) on latency at higher bandwidths (≥50 Mbps). These findings are complement to the RQ4 and RQ5 findings.

Figure 13 shows the line graphs of the measures collected for latency vs bandwidth of Mobile-Edge (Partitioned, Quantized Early Exit Partitioned), Mobile/Edge (Identity, Early Exit, Quantized, and Quantized Early Exit) deployment strategies. For DUC (orange line), the Mobile-Edge Partitioned-based models show a sharp decrease in latency when Mobile-Edge bandwidth increases from 1 Mbps to 50 Mbps. In the same case, the Mobile/Edge (Identity, Early Exit, Quantized, Quantized Early Exit) deployment strategies show a marginal drop in latency. The reason is the higher size of intermediate data (781.25Mb) compared to the input data (22 Mb) of DUC and that's why the latency of Partitioned-based models across Mobile-Edge shows a sharp decrease as the bandwidth increases from 1 to 50 Mbps. For the other subjects, i.e., FCN (green line), ResNet (purple line), and ResNext (blue line), the intermediate data is not that significantly large compared to DUC; therefore, a smaller decrease in latency is shown from 1 to 50 Mbps. For ResNe(x)t, the input data size is the highest (3x to 12x) compared to the other subjects (FCN, DUC), which is why it shows a larger decrease in latency for the same interval when models are deployed on Mobile and Edge tiers. For FCN, the trend is overall quite steady for the models during Mobile and Edge deployment as the input data size is quite small (5 Mb) to make an impact on the latency for various network bandwidths. These findings at low-bandwidth scenarios of Mobile-Edge

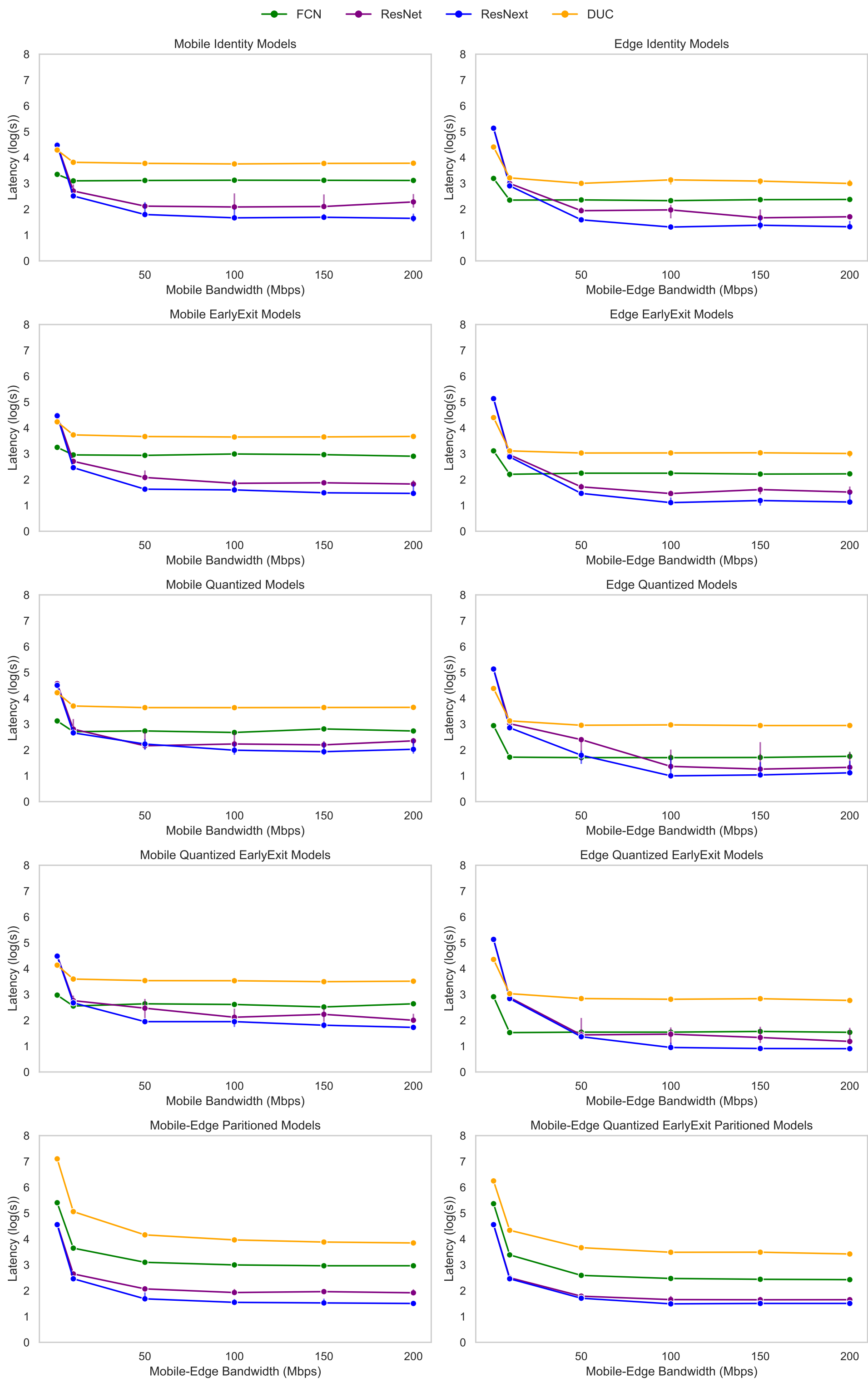


Fig. 13: Line graphs of the measures collected for inference latency vs network bandwidth of Mobile-Edge Partitioned, Mobile-Edge Quantized Early Exit Partitioned, Mobile (Identity, Early Exit, Quantized, Quantized Early Exit), and Edge (Identity, Early Exit, Quantized, Quantized Early Exit).

($\leq$50 Mbps), complement the RQ1-5 findings of high Mobile-Edge bandwidth (i.e., 200 Mbps). Table 24 shows a summary of results for RQ6.

5.6.2 Normality Test

We conducted the Shapiro-Wilk normality test and examined QQ plots for each model variant and bandwidth configuration. Contrary to typical expectations, the Shapiro-Wilk test results indicated that the latency distributions for the majority of configurations did not significantly deviate from normality ($p > 0.05$), as shown in Table 31, Table 32, Table 33, Table 34, Table 35, Table 36, Table 37, Table 38, Table 48, Table 46, Table 47, Table 39, Table 40, Table 43, Table 44, Table 41, Table 42, and Table 45, Figure 33, Figure 34, Figure 35, Figure 36, Figure 37, Figure 38, Figure 39, Figure 40, Figure 41, and Figure 42. This suggests that latency data maintains a roughly symmetric distribution with minimal skewness or outlier influence across diverse scenarios. These findings were visually corroborated by the QQ plots, which showed points aligning closely with the reference line.

5.6.3 Hypothesis Testing

As shown from Table 49 to Table 154, for all model configurations, the Kruskal-Wallis p-value is extremely small ($p < 0.05$), which indicates strong evidence against the null hypothesis. This suggests that there are significant differences in latency across the different bandwidth conditions for each model. After finding significance with the Kruskal-Wallis test, Conover's pairwise post-hoc tests were used to pinpoint which specific pairs of bandwidth values differ significantly when Mobile-Edge bandwidth is fixed and Edge-Cloud bandwidth is varied, and when Edge-Cloud bandwidth is fixed and Mobile-Edge bandwidth is varied. In both scenarios, Conover post hoc tests showed that models like ResNet, ResNeXt, FCN, and DUC exhibited strong sensitivity to changes in the Edge-Cloud and Mobile-Edge bandwidths, with most low vs. high bandwidth comparisons producing significant p-values (typically $p < 0.05$). In general, the findings indicate that both the ME and the EC bandwidths significantly influence the latency of the model, especially at lower bandwidth levels, while the performance differences diminish at higher bandwidths (e.g., 100–200 Mbps), suggesting a saturation effect.

Summary of Research Question 6

Deploying models with small input data size (i.e., FCN) is ideal for Cloud deployment with bandwidth-constrained settings ($\leq$10 Mbps). ResNe(x)t and DUC models with large input data size are suitable for Cloud deployment that operate in scenarios with at least moderate bandwidth availability ($\geq$50 Mbps) but might not be ideal for constrained bandwidth settings ($\leq$10 Mbps). For models with lower intermediate data size (i.e. ResNe(x)t), the Edge-Cloud/Mobile-Cloud Partitioned-based strategies have no impact of intermediate data on latency across network bandwidth variations.
For models with higher intermediate data size (i.e., FCN, DUC), the Edge-Cloud/Mobile-Cloud Partitioned-based strategies either converge at higher bandwidths ($\geq$100 Mbps for FCN) or doesn't converge (for DUC). For Mobile-Edge Partitioned based strategies, all the subjects converge at 50 Mbps (except DUC, which converges at 100 Mbps due to higher intermediate data size). For ResNe(x)t, the input data is the largest (3 to 12x) compared to the other subjects (FCN, DUC), which is why it shows a larger decrease in latency at 50 Mbps for Mobile and Edge deployments before plateauing.

6 Discussion

6.1 Interpretation of Results

Among the three Monolithic deployment tiers, the Edge tier consistently demonstrates significantly faster inference latency performance for each operator examined in the respective research questions

Model	Ideal Bandwidth (Mbps)	Best Tier	Latency Behavior	Notes
ResNet	≥ 50	Edge/Cloud	High drop from 1–50 Mbps	Input size (60MB) leads to latency reduction as bandwidth improves
ResNext	≥ 50	Edge/Cloud	Large decrease, then plateaus	Cardinality increases memory usage; sensitive to bandwidth change
FCN	$\leq$10	Cloud	Stable across bandwidths	Small input/intermediate size makes it bandwidth-resilient
DUC	$\geq$100	Edge	Sharp drop, then plateau	Huge intermediate size (781MB); converges at 100 Mbps

Table 25: Summary of results for Research Question 6

(i.e., Identity operator in RQ1, Quantized operator in RQ2, Early Exit operator in RQ3, and Quantized Early Exit operator in RQ5). This outcome underscores the significance of deployment operators involved in the Edge tier compared to Mobile and Cloud tiers when facing computational limitations, network bandwidth constraints, and large input data transmission. The Edge tier's closer proximity to the Mobile and higher computational resources allows faster processing and reduced transmission latency, resulting in improved inference latency performance.

In the Cloud tier, the comparisons between the Quantized operator and the Identity operator (RQ2) as well as between the Quantized Early Exit operator and the Early Exit/Identity operator (RQ5) do not show any significant difference (Conover Test) for any subjects, except DUC. The main challenge is the inability of the CUDA execution provider in the ONNX Runtime inference Engine to fully support the CUDA kernels used for the quantized graphical nodes during Cloud deployment. The unavailability of the CUDA kernel for the nodes of the Quantized/Quantized Early Exit models leads to their execution being run on the CPU instead. This introduces some overhead due to the lower processing power of the CPU compared to the CUDA. Future research efforts can delve into solutions for optimizing the Quantization in a GPU-based Cloud environment. One solution might be to modify the model structure and its operations to avoid nodes that lack CUDA kernels. Techniques such as operator fusion, where multiple operations are combined into a single operation, could be beneficial [8]. Another solution could be to develop custom CUDA kernels for the nodes of the Quantized/Quantized Early Exit models that currently lack them. This would require a deep understanding of both CUDA programming and the specific operations performed by the incompatible nodes. Utilizing other GPU-specific execution providers like TensorRT instead of CUDA was not considered due to the reasoning provided in Section 7.

On Cloud, the comparisons between the Early Exit and Identity operator (RQ3) as well as between the Quantized Early Exit and Quantized operator (RQ5) show no significant difference for the Resnet/ResNext/FCN subject and even if the difference is significant (in the case of DUC) based on the post-hoc test, the effect size remains negligible to small. This suggests that the advantages provided by these specialized operators in terms of speedup may be less pronounced in Cloud deployments, where computational resources are typically more abundant. In the above scenarios, the use of the Identity operator alone may be sufficient, and incorporating specialized operators like Quantization, Early Exit or their combinations may not yield significant benefits in terms of improving latency.

Among the three multi-deployment tiers (Mobile-Edge, Edge-Cloud, and Mobile-Cloud), the Mobile-Edge tier consistently exhibits faster latency performance for each operator examined in their respective research questions (i.e., Partition operator in RQ4 and Quantized Early Exit operator in RQ5). The key contributing factor to the Mobile-Edge tier's superiority is the higher network bandwidth it offers, which facilitates faster transmission of intermediate outputs during distributed inference. The Mobile-Edge distributed inference consistently shows faster latency performance than stand-alone Mobile inference for Identity models in RQ4 and Quantized Early Exit models in RQ5 due to the computational load distribution. The Partitioning of Identity/Quantized Early Exit models should ideally not influence the accuracy drops as it aims to divide the model into smaller components without altering the computations or operations performed, as stated in [74]. In other words, the computations within the Partitioned models are consistent with the original model's computations. The models considered in previous studies for CV tasks are relatively small (i.e, lower size) and less complex, which potentially resulted in findings that Mobile deployment is a better alternative than Mobile-Edge Partitioning (Kang et al [50]). However, in our study, where complex models are considered as the subjects, Partitioning across the Mobile and Edge is a better alternative than doing the local computing of the whole model on a resource-constrained Mobile tier, especially when faster latency is a concern at no accuracy loss.

This study considers multiple dimensions, including operators, data models, network bandwidth, and tiers, requiring us to control one parameter at a time to analyze its impact on the others. In RQ1–RQ5, we fixed the network bandwidth at 1 Mbps for Mobile-Edge and 200 Mbps for Edge-Cloud to examine the effects of different operators and data models. This choice was deliberate, as these values are commonly used (as discussed in Section 4), whereas RQ6 explored a range of bandwidths to assess how variations influence latency performance. The results from RQ1–RQ5 indicate that operators (Identity, Quantized, Early Exit, and Quantized Early Exit) exhibit better latency performance on Edge compared to Cloud, primarily due to the Edge tier's higher bandwidth (200 Mbps vs. 1 Mbps) and the impact of input data transmission size. In RQ6, a similar relative performance was observed when Mobile-Edge and Edge-Cloud bandwidths were $\leq$ 50 Mbps; however, when bandwidth exceeded 50 Mbps, the Cloud tier outperformed the Edge tier. Furthermore, RQ6 results show that Mobile-Edge Partitioning-based strategies consistently outperform both Mobile-Cloud and Edge-Cloud Partitioning-based strategies, reinforcing findings from RQ4 and RQ5. Additionally, Mobile-Edge Quantized Early Exit Partitioning-based models demonstrate superior latency performance compared to Mobile-Edge Partitioned models, further confirming that bandwidth constraints significantly impact partitioned strategies, as previously observed in RQ4 and RQ5.

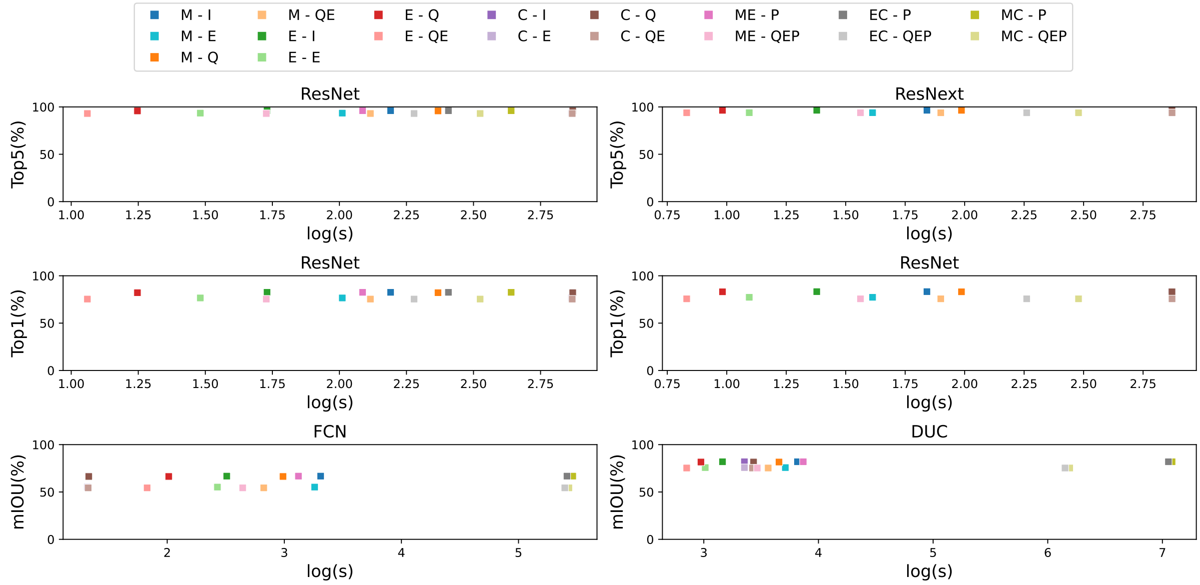
We examine how network bandwidth variations, explored in RQ6, influence the outcomes of RQ1–RQ5. Under the constrained 1 Mbps Cloud bandwidth assumed in RQ1–RQ5, the Edge tier consistently provides superior latency performance (RQ1–RQ3). However, when the Cloud bandwidth increases to 50 Mbps or more (RQ6), the Cloud tier becomes the optimal choice, emphasizing the significant role of network capacity. In RQ4, Mobile-Edge partitioning is most efficient under the 1 Mbps Cloud bandwidth, but at higher bandwidths, Edge-Cloud partitioning outperforms it by combining fast data transfer with Cloud computation. In RQ5, QE on Edge and QEP on Mobile-Edge perform best under low Cloud bandwidth, whereas at >50 Mbps, QE on Cloud emerges as the fastest monolithic strategy, and QEP on Edge-Cloud becomes the most effective hybrid strategy, surpassing Mobile-Edge setups. Mobile-Edge bandwidth also plays a critical role in the effectiveness of partitioned strategies such as Mobile-Edge Partitioned and Mobile-Edge QEP. While these strategies perform well under the 200 Mbps assumption used in RQ1–RQ5, their performance deteriorates when the Mobile-Edge bandwidth is reduced (e.g., 1–50 Mbps), as the cost of transmitting large intermediate data offsets the benefits of distributed inference. Likewise, models such as DUC, which generate large intermediate outputs, show poor performance under constrained Mobile and Mobile-Edge bandwidth conditions, but perform significantly better when bandwidth exceeds 100 Mbps.

In the accuracy vs latency scattered plots (Figure 14) of the deployment strategies when evaluated on a range of input samples, we can see that for ResNet, ResNext, and DUC, the Edge Quantized Early Exit (E - Quantized Early Exit) deployment strategy (light red scattered points) is Pareto-dominated by other strategies in terms of latency. Whereas, for FCN, the Cloud Quantized (C - Q) deployment strategy (dark brown scatter point) is Pareto-dominated by other strategies in terms of all objectives (i.e., latency and accuracy).

The significant presence of outliers in the box plots of ResNext on Edge (Figure 4) is possibly due to the higher memory usage while achieving inference latency measurements on Edge. This is primarily due to its cardinality-based architecture, which provides better accuracy than Resnet at the cost of increased computational and memory requirements [9]. Here is a detailed breakdown of the outliers in the latency results of ResNext on Edge:

1. Cardinality in ResNext: ResNeXt introduces the concept of cardinality, which refers to the number of parallel paths or branches in a block. Each ResNeXt block splits the input into multiple branches, processes them independently, and then aggregates the results. While this improves representational power and flexibility, it increases:
   - Intermediate activations: Each branch produces its own intermediate feature maps, increasing the total memory required to store these activations during forward and backward passes.
   - Parameter storage: Each branch has its own convolutional layers, leading to a greater number of weights to store in memory. For larger cardinalities, more parameters are distributed across branches. This increases the memory required to store weights and biases.
2. Aggregation of Branch Outputs

[9] https://www.ikomia.ai/blog/resnext-cnn-cardinality-efficiency-explained



M-I: Mobile Identity, M-E: Mobile Early Exit, M-Q: Mobile Quantized, M-Quantized Early Exit: Mobile Quantized Early Exit, E - I: Edge Identity, E- E: Edge Early Exit, E-Q: Edge Quantized, E - Quantized Early Exit: Edge Quantized Early Exit, C - I: Cloud Identity, C - E: Cloud Early Exit, C - Q: Cloud Quantized, C - Quantized Early Exit: Cloud Quantized Early Exit, ME - P: Mobile Edge Partitioned, ME - Quantized Early Exit: Mobile-Edge Quantized Early Exit Partitioned, EC - P: Edge Cloud Partitioned, EC - Quantized Early Exit: Edge Cloud Quantized Early Exit Partitioned, MC - P: Mobile Cloud Partitioned, MC - Quantized Early Exit: Mobile Cloud Quantized Early Exit Partitioned.

Fig. 14: Scattered plots of the measures collected for latency vs accuracy of the deployment strategies when evaluated on a range of input samples.

   – After processing through the branches, ResNeXt aggregates their outputs (usually by summation). This process temporarily requires additional memory to store the outputs of all branches before combining them.
   – In ResNet, this step is simpler since there is only a single path per block, avoiding this additional memory overhead.
3. Wider Representations: ResNeXt achieves greater representational power by increasing cardinality rather than depth or width. While this improves accuracy, it also:
   – Increases the size of intermediate tensors: Wider representations mean larger activation maps, which consume more memory.
   – Requires storing gradients: During back-propagation, the gradients of these wider activations must also be kept in memory, further increasing the memory demand.
4. Redundant Memory Usage in Backpropagation: During training, intermediate activations are stored for gradient computation. In ResNext:
   – Each branch has its own set of activations that must be retained.
   – The memory required for these intermediate results grows linearly with the cardinality.
5. Suboptimal Hardware Utilization: Many hardware accelerators are optimized for simpler, sequential architectures like ResNet. The parallel branch design in ResNext can lead to inefficiencies in memory allocation and access, indirectly contributing to higher memory usage.

These outliers, which represent higher inference latencies, narrow the gap between the Edge and Cloud latency of ResNext, thereby reducing the perceived advantage of Edge in specific scenarios. In particular, these outliers on Edge indicate that under memory-constrained conditions caused by ResNext cardinality-based architecture, Edge may experience occasional latency spikes, which could undermine its advantage over Cloud. The majority of data points and the median on Edge tightly cluster below Cloud's latency median, emphasizing the general trend of lower inference on Edge. For latency comparison, as we considered the median, which itself is outlier-insensitive, the presence of outliers does not necessarily affect the overall conclusions. This suggests that Edge indeed is a preferable choice over Cloud for

ResNext. Moreover, while comparing the distribution of Edge and Cloud latency of Resnext, we see a large effect size (Table 8) indicating a statistically significant difference.

6.2 Comparison with Existing Literature

In previous studies (Table 1), the resolution of input data size is relatively small as they are validated only on small-scale datasets such as MNIST and CIFAR datasets, and/or they did not explicitly analyze the impact of high-resolution images (from Image Net, COCO, and CityScapes datasets) on the end-to-end latency evaluation of operators in Mobile, Edge, and Cloud tiers. Eshratifar et al. [24] suggest that using either local computing only or Cloud computing only is not an optimal solution in terms of inference latency in comparison to model Partitioning. However, the key issue is that they considered a single image for the end-to-end sequential inference across the Mobile-Cloud tier, without exploring the variation of input data sizes. In our study, the impact of multiple and varying image sizes on end-to-end latency was explored, which resulted in more generalized findings. Based on our results, the subjects having low-resolution images (such as FCN) may favor network-constrained Cloud deployment in comparison to multi-tier Partitioning strategies and Mobile/Edge deployment, as the impact on transmission overhead for smaller-sized images reduces, and the Cloud, as usual, has better computational capabilities. For subjects (i.e., ResNet, ResNext, DUC) having large-sized image samples, Edge deployment is a better alternative than multi-tier Partitioning strategies and Mobile/Cloud deployment.

Prior work [24,48,50,59,123,88,20,120,80,62,43] explores factors such as the computational load, network cost, energy consumption, and/or privacy risk for each of the DNN Partitioning points in an Edge AI setup to dynamically decide the optimal Partition point, and stated that the model Partitioning operator achieves significant latency speedup i.e., latency reduction compared to traditional Mobile and/or Cloud deployment, similar to our findings, i.e., Mobile-Edge distributed inference of Partitioned/Quantized Early Exit models is a better alternative than resource-constrained Mobile deployment of Identity/Quantized Early Exit models. When comparing the multi-tier distributed strategies of Partitioned/Quantized Early Exit models with the Cloud deployment of Identity/Quantized Early Exit models in RQ4/RQ5, the intermediate data size and input data size play a crucial role. For FCN and DUC, the intermediate data size of their Partitioned/Quantized Early Exit variants is larger than their input data sizes, correlating with faster Cloud inference than distributed inference. Conversely, for ResNet and ResNext, the intermediate data size of their Partitioned/Quantized Early Exit variants is smaller than their input data sizes, correlating with faster-distributed inference than Cloud inference. Different Partition points (specific graphical node connections (s) within the neural network architecture where the model is divided or split into two sub-models) might have different intermediate data sizes, and their impact on the latency might vary. However, in our study, we limited our research to a single Partition point, as the goal was to create equal-size Partitioned models (which require a single Partition point). Although our Partitioning approach simply and fairly splits the models statically into two sub-models to have equal sizes, one running at the Mobile/Edge, and the other one in the Edge/Cloud, it requires manual analysis of the ONNX computational graphs of the subject models, which varies in terms of graph complexity and architecture design. In terms of subject models considered for model Partitioning, previous studies performed their experiments on lightweight CV models, which are less complex and less accurate than the heavy-weight state-of-the-art CV models considered in our study.

In previous studies performing Early Exit [23,54,66,91,102,106,107,108,111,116,118,119,123,128, 65,73,117,113], there is a trade-off between accuracy and latency. This trade-off in Early Exit comes from the fact that exiting earlier in the network can reduce latency but may also result in less accurate predictions. This is because the early layers in a DNN generally extract low-level features, while the later layers extract high-level features that are more task-specific. Therefore, an early exit at early layers might miss important high-level features, leading to a decrease in accuracy. On the other hand, waiting for the network to reach the later layers can increase the accuracy but also increase the latency. In our case, the Early Exit was performed in the later stage of the models, which showed faster latency than the original model but at a medium accuracy loss. One of the reasons for this significant accuracy loss is that the Early Exit approach in our study is based on the condition of manually short-circuiting identically structured sub-graphs on pre-trained models. This means that the early exits are added in a black-box manner on pre-trained models without retraining them, which contributes to this accuracy loss. Previous studies perform DNN Early Exit that requires training of the models, which results in better accuracy performance for Early Exit even in earlier stages of the model.

As stated in previous studies [59, 5, 7, 12, 27, 29, 40, 57, 77, 83, 101, 30, 45, 126, 61], the Quantization operator in our study also shows a small (not significant) accuracy drop in comparison to the original model. In addition to that, we also compared the Quantization operator's performance with other operators like Early Exit and Quantized Early Exit operators, concluding that during Edge deployment, the Quantization can be used in scenarios where the least accuracy drop is of utmost importance w.r.t the original model, at the benefit of faster latency than Early Exit and the cost of slower latency than Quantized Early Exit operators.

6.3 Implications for the Practitioners

In terms of effort, the use of an automated tool like Intel Neural Compressor for applying the Quantization operator suggests a streamlined and automated process for Edge AI model deployment. This implies that, right now, the Quantization process can be performed without manual effort and intervention, as a tool automates the necessary modifications to achieve model Quantization. Yet, while applying Quantization to models from the ONNX Model ZOO and torchvision.models subpackage, we observed that some of the models contain unsupported ONNX operations that are not listed in the supported ONNX schema[10], due to which the model may not be feasible for Quantization without additional modifications or workarounds. For instance, while converting models from Pytorch to ONNX format, the converted ONNX models can include custom layers or operations that are specific to the Pytorch framework and hence lack Quantization support in ONNX.

On the other hand, applying the Early Exit operator on Identity/Quantized models and the Partitioning operator on Identity/Quantized Early Exit models currently requires manual analysis of the ONNX computational graphs using the Netron Visualizer tool [11] and manual modifications of the neural network using ONNX Python APIs. We observed that performing these deployment operators on ONNX models from other classes (e.g., textual inference task), was not always feasible due to the complex architecture of the ONNX computational graph. The complex architecture of such models includes various layers, connections, and branching structures that can make it difficult to identify suitable Partition points or early exit points. In such cases, applying these operators might require significant manual analysis and modifications of the model's computational graph, which can be time-consuming and error-prone. Below are the crucial points that highlights why the results are helpful for MLOps and in which cases:

Empirical Foundation for Deployment Choices The study delivers data-backed insights on how different black-box deployment operators (Quantization, Early Exit, Partitioning) and their combinations affect latency and accuracy across Mobile, Edge, and Cloud tiers. MLOps engineers often struggle with trial-and-error deployment tuning—this study systematically removes the guesswork by evaluating 20+ configurations in controlled, realistic environments.

Our study offers a concrete empirical foundation to assist MLOps engineers in selecting appropriate deployment strategies for black-box models across heterogeneous Edge AI environments. For instance, when aiming to reduce latency under moderate accuracy constraints, Quantized Early Exit (QE) on Edge emerges as a promising solution due to its effective balance between performance and computational cost. Conversely, when preserving accuracy is critical, using Quantization alone on Edge is preferable, as it delivers latency benefits with minimal accuracy degradation compared to Early Exit or QE. In resource-constrained Mobile environments, Mobile-Edge Partitioning outperforms pure Mobile deployment by offloading heavy computation to nearby Edge tiers, reducing latency without sacrificing output fidelity. Additionally, the results show that Cloud deployment remains viable for smaller input models even under low-bandwidth conditions ($\leq$10 Mbps), while larger models require at least 50 Mbps for latency convergence. These insights can help MLOps teams systematically evaluate the trade-offs between latency, accuracy, and resource constraints, avoiding ad hoc trial-and-error tuning and promoting informed, performance-aware deployment strategies.

Black-Box Compatibility The operators analyzed require no model re-training, which is ideal for real-world, production-grade MLOps settings where models are often closed-source or externally sourced. This makes the findings highly actionable across industries.

[10] https://github.com/onnx/onnx/blob/main/docs/Operators.md
[11] https://github.com/lutzroeder/netron

Towards a Foundational Dataset for Recommendation Systems The empirical results of the study provide valuable insights into the latency and accuracy performance of operators across and within tiers under various deployment scenarios. As such, we believe that the results could inspire work related to the automation of these operators and can be used as a foundation for developing recommendation systems for Edge AI operators. In particular, the results serve as core training data for automated systems such as AGI-based recommendation engines that learn optimal AI deployment strategies. This includes input-output mappings for a wide range of models, operators, tiers, and network conditions—allowing autonomous adaptation in real-time environments.

However, we recognize that our current insights are derived from a simulated and controlled testbed, which may not capture all real-world heterogeneities (e.g., hardware variation, dynamic workloads). As such, we do not position our results as directly prescriptive but rather as a knowledge base from which data-driven heuristics could eventually be extracted and validated in real environments. To ensure trust and transparency, any future recommendation system must be accompanied by interpretable explanations, and its performance must be benchmarked against real-world deployment scenarios. Until such validation is achieved, we encourage practitioners to rely on the detailed experimental conditions we provide in this study to assess contextual relevance to their own settings.

While this study focuses primarily on latency and accuracy trade-offs, our observations about memory-boundedness suggest the importance of tracking CPU utilization, memory footprint, network throughput, and energy consumption. These system-level metrics are highly relevant for environmentally-conscious deployments and for fine-grained performance optimization. Future work will expand the current benchmarking pipeline to include these metrics, which are already partially accessible via Docker container statistics and external energy profiling tools (e.g., Intel RAPL, NVIDIA SMI). Doing so will allow us to further align Edge AI deployment strategies with sustainability goals.

In the Future, we plan to extend our testbed with tools like cgroups, perf, and powerstat to collect metrics such as CPU usage, RAM consumption, and energy draw during inference. This will support a multi-objective evaluation of deployment strategies, including carbon footprint and energy cost, critical variables in real-world AI systems design.

Versatile B2B Applicability These results are not just academic—they can be directly applied in any B2B use case where AI models are deployed in an Edge AI Infrastructure (e.g., manufacturing, healthcare, smart cities, autonomous vehicles, and more).

7 Threats to Validity

Below, we discuss threats to the study validity and the strategies we applied to mitigate these threats, based on literature guidelines [112].

Construct Validity: One possible threat is the mono-operation bias caused by having only one factor of computational configuration (RAM/CPU), Early Exiting point, and the Partitioning point. The Mobile-Edge network bandwidth of 200 Mbps and Edge-Cloud network bandwidth of 1 Mbps were used to simulate the close and distant proximity of Mobile-Edge and Edge-Cloud environment, respectively, based on earlier work [84,125,103,28,3] for testing deployment strategies on a range of samples in RQ1-5. We further expand our analysis to include additional experiments evaluating deployment strategies on a single largest input sample under a broader range of commonly used bandwidth settings (1, 10, 50, 100, 150, and 200 Mbps) across both Mobile-Edge and Edge-Cloud networks [1,17,130,17,84,125] in RQ6. The additional findings in RQ6 results (Section 5.6) allow us to provide a more comprehensive understanding of the latency performance trade-offs under varying bandwidth conditions and emphasize the impact of bandwidth on deployment strategies. The computational simulations of Mobile, Edge, and Cloud tiers are also based on previous studies [19,21,53,94]. Variations in the computational resources can impact the end-to-end inference latency for the deployment strategies.

One of the key limitations of our study is the use of Docker containers for simulating resource-constrained tiers (Mobile, Edge) on a common server and resource-abundant tier (Cloud) on a separate server in our experimental setup. We recognize that a distributed setup using multiple physical machines would provide additional realism. Our Docker-based simulated experimental setup was designed to ensure the reproducibility of our results. While we acknowledge that this approach may not fully replicate the complexities and challenges faced in real-world distributed MLOps deployment infrastructure, our primary objective was to provide a controlled and consistent environment for testing and validating our deployment strategies. This methodology aligns with prior research, as highlighted in Table 1 of the Related Work section (Section 3), where similar simulated setups have been employed. By adopting this approach, we establish a reproducible framework that can serve as a foundation for future studies. We further test a comprehensive range of network bandwidths ensuring our evaluation aligns with the diverse characteristics of real-world deployments.

However, our decision to use Docker containers on a single machine was intentional and driven by the following considerations:

- Prototyping a Unified Baseline for MLOps: MLOps engineers often start by testing their pipelines in controlled environments to ensure baseline functionality before scaling to distributed systems. Our choice mirrors this natural progression, ensuring that our findings are directly relevant to the foundational stages of MLOps pipeline development.
- Reproducibility as a Scientific Priority: One of the key challenges in evaluating MLOps tools and strategies is achieving reproducibility across studies. A single-machine Docker-based setup provides a stable and standardized environment, reducing variability and enabling others to replicate our work with minimal dependencies or hardware constraints. This step is crucial in building trust and validating methods within the community.
- Cost-Effective Innovation: MLOps research is inherently resource-intensive. By leveraging a Docker-based approach, we reduced the financial and logistical overhead of deploying experiments on distributed physical systems. This resource efficiency allowed us to focus on developing innovative insights into key MLOps challenges, with the understanding that future work can build on these foundations in more distributed contexts.
- Alignment with MLOps Practices: Docker-based setups are widely used in MLOps workflows, particularly during the prototyping phase (Table 1 in Related Work Section). Our approach aligns with this practice, ensuring the immediate applicability of our results to real-world scenarios.

The limitations of the considered simulated setup:

- Clock Speed and Thermal Factors: Docker provides a valuable feature for setting resource constraints, enabling the simulation of specific configurations like the number of CPU cores and available RAM. To further strengthen the study, it's worth noting that certain hardware-level characteristics, such as clock speed variations and thermal management, may differ from real-world systems.
- GPU Utilization and Variability: The study's cloud simulation utilizes an NVIDIA A100 GPU, which is a state-of-the-art piece of hardware. However, to enhance the study's applicability to real-world scenarios, it might be beneficial to consider the diversity of GPUs typically deployed in actual cloud environments, such as the T4, V100, or even older hardware. This could provide a more comprehensive understanding of performance across different cloud setups.

Influence of these limitations on the study findings:

- Limited Real-World Representation: The results might not fully capture the nuances of actual hardware behaviors, such as clock speed variations or thermal management, which are significant in real-world settings.
- Performance Variability: Differences in computational and hardware characteristics may lead to outcomes that do not align perfectly with what would be observed on diverse real-time systems.

In a previous study [93], Docker containers running on a server are used to simulate several resource-constrained tiers in a realistic IOT framework. This is similar to what we did for simulating the resource-constrained tiers (i.e., Mobile and Edge) in our study. In real-world scenarios, variations in hardware across different Edge/Mobile devices are possible and may impact the generalizability of our simulated setup. Simulating the impact of multiple factors is very costly as it involves running numerous experiments, each taking considerable time to complete and requiring significant computational resources. While we acknowledge the importance of diverse deployment scenarios having different computational/network configurations of Mobile, Edge, and Cloud tiers, the specific experimental setup was chosen strategically to provide a focused exploration of a typical Edge AI environment having a resource-scarce Mobile device, an Edge device's closer proximity, and higher computational capacity w.r.t Mobile device, together with a resource-abundant Cloud device with network constraints.

In our study, we focus on Early Exit at a single stage of the neural network. However, it is important to note that early exiting at multiple stages of the network can result in varying accuracy performance.

Early exiting at a later stage may yield higher accuracy but slower inference due to more processing required before making predictions. On the other hand, the early exit at an earlier stage may provide faster inference but with lower accuracy since the predictions are made based on less processed information. The detailed explanation for selecting the early exit criteria is mentioned in the RQ3 approach (Section 3.4), and requires manual inspection of the subjects' ONNX computational graphs. Similarly, for Partitioning, we considered the Partition point that leads to equal-sized sub-models for effective and fair load distribution across the tiers involved in distributed inference. Our study acknowledges the limitations associated with this simplified model Partitioning approach. We recognize that real-world scenarios often entail more complex Partitioning strategies as mentioned in related work (Section 1), especially when dealing with intricate model architectures, such as Partitioning at each of the layers of DNN models and analyzing their impact on various factors like computational, transmission, and/or energy cost.

In our study, the decision to adopt a simplified approach stems from the need for a fair evaluation of the subject models. By manually inspecting the ONNX computational graphs of the DNN models, we aim to establish a baseline understanding of the challenges and dynamics involved in equal-size Partitioning of the DNN models with varying and complex architectures, specifically in the ONNX framework. The manual insights of the computational graphs of the subjects will assist MLOps Engineers in understanding the factors involved in constructing an automated tool to dynamically decide the optimal Partition point to achieve equal-size sub-models for subjects with varying architectures. This approach might have an adverse impact on the transmission of intermediate data during distributed inference across the Mobile, Edge, and Cloud tiers, as shown for two of the subjects (FCN, DUC) in our study. Different Partition points may have different intermediate data sizes which can yield latency benefits during distributed inference across Mobile, Edge, and Cloud tiers in scenarios where the intermediate data size is lower than the input data size, as shown (ResNet and ResNext) in our study.

Conclusion Validity We considered the Static PTQ approach over the Dynamic PTQ approach due to its faster inference capabilities. Dynamic PTQ requires additional computational overhead during inference because of the dynamic recalibration process in which the model's weights and activations are recalibrated based on the input data's statistics during inference. On the other hand, Static PTQ involves quantizing the model's weights and activations only once during the model conversion phase, without the need for recalibration during inference. Since the quantization parameters are precomputed and do not change during inference, the quantization process is much simpler and requires fewer computations during inference.

Internal Validity: The risk of how history might affect the inference latency results of the deployment strategies is reduced by performing all measurements in the same Edge AI environment, using the same infrastructure. To maintain uniformity and minimize variations, we developed automated scripts to execute the inference experiments for the deployment strategies sequentially, one after another. Before starting an inference experiment, we took the necessary step of restarting the Docker containers to eliminate any potential residual effects from the previous inference experiment. In our experiments, we considered sequential inference (i.e., 1 request at a time) instead of parallel inference (i.e., multiple requests at a time). Sequential inference allows us to efficiently utilize the available resources for each deployment strategy, ensuring a more accurate representation of their true inference latency performance [124], similar to how micro-benchmarks operate. The goal of our study was not load testing [49], where the focus would be on measuring the system's ability to handle multiple concurrent inference requests. Performing parallel inference may lead to resource contention, which could obscure the true impact of deployment strategies. Similarly, the Scalability aspects, such as the impact of increasing the number or complexity of models deployed simultaneously in a real-world setting, are not explored, as they are outside the scope of this study, since our micro-benchmarks will not focus on system-level measurements.

It is worth mentioning that the ONNX Run-time inference Engine may perform worse on the first input received than on subsequent inputs during inference experiments of deployment strategies, mainly because of a required warm-up inference. To remove such bias, we used a trial inference experiment. For each deployment strategy, a trial experiment of 100 inference runs is performed sequentially to reach a steady state of the cache, then the final inference experiment of 500 runs = 100 (input samples) x 5 (repetitions) is performed sequentially (and repeatedly) without any cool-down period between subsequent runs to simulate the scalability of each deployment strategy. Note that the relatively high standard deviation in the inference latency measure for each deployment strategy might have been caused by having 5 repetitions per sample run; this potential source of bias can be mitigated by increasing the number of such repetitions. In our study, it is costly to do this due to the computational and transmission overhead caused by factors like large model size and input/intermediate data transmission.

Input size is a potential threat to internal validity, as it can significantly affect both latency and resource usage across models. To address this, we carefully selected 100 large-size input samples from the validation dataset for the multi-input trial experiments. This selection aimed to evaluate the impact of varying input sizes on latency, as described in Section 4.6. Additionally, for the single-input trial experiments, we used the largest input sample available to assess latency across different deployment strategies and network conditions. Testing with larger input sizes allowed us to evaluate how well the simulated environment can handle heavier computational workloads, simulating real-world scenarios such as medical imaging, video games, and high-resolution photography. This approach also helped assess the scalability of the system and identify potential bottlenecks, such as increased latency due to memory, processing power, or bandwidth limitations. By carefully considering and controlling for input size in our experimental setup, we aimed to minimize its potential confounding effects on the results.

External Validity: The selection of the subjects might constitute another potential threat as it was performed manually. Thus, the selected set of subjects could not be regarded as an accurate representation of the whole population. The first iteration in this process consisted of choosing a set of subjects from the ONNX Zoo and PyTorch Models, initially aiming to have two representatives from each class (i.e., the inference task). However, we observed a lack of already trained models from some of the classes and what is more, most of the models were not feasible for Partitioning/Early Exiting due to the complex ONNX graph architectures, or for Quantization due to some operations or layers in the ONNX graph architectures that may not have Quantization support in the neural compressor tool. As a result of that, we ended up with a selection of four subject models, from the CV category. This threat is reduced by aiming to diversify the inference tasks they performed (i.e., Image Classification and Image Segmentation).

The majority of the previous studies, as shown in Table 1, focus on CV tasks for the operators due to their major impact on various factors like computational load and data transmission during deployment in an Edge AI environment. Therefore, the choice of CV domain for subject models and datasets is quite common. However, we do acknowledge that different types of models (e.g., Natural Language Processing and Speech Recognition) may exhibit different behaviors in response to deployment strategies. In our study, we limited the scope of our study to three black-box operators, i.e., Partition, Quantization, and Early Exiting, as these are most commonly used, which do limit the comprehensiveness of the studied operators. Overall, the selection of subjects (datasets/models) and deployment operators is one of the external threats concerning this study. This threat can be mitigated in the future by repeating the experiment on other domain-specific subjects (e.g. Natural Language Processing, Speech Recognition) and white-box operators (e.g., Quantization Aware Training, Weight Pruning, Knowledge distillation) as mentioned in Table 1.

The effectiveness and compatibility of the Intel Neural Compressor tool for Quantization might vary for different models or frameworks (i.e., ONNX), affecting the reproducibility of the study in different environments (i.e., CPU, GPU). In our results, for some subjects like ResNet and ResNext, their Quantized and Quantized Early Exit models show slower latency performance than Identity and Early Exit models in resource-constrained environments (i.e., Mobile). Yet, for other subjects (like FCN and DUC), their Quantized and Quantized Early Exit models show faster latency performance than Identity and Early Exit models in the same environment. Conversely, in high-resource environments (i.e., Edge), all subjects' Quantized and Quantized Early Exit models show faster latency than the Identity and Early Exit models. This shows that for different subjects the compatibility of this tool (i.e., Intel Neural Compressor) for Quantization might vary in terms of latency performance in computationally varying environments. Moreover, this tool might show different latency or accuracy behavior for models in different formats (such as Pytorch and Tensorflow). Moreover, the visual analysis of the computational graphs might vary with different models, frameworks, or visualization tools.

Among the GPU-specific Execution Providers (e.g., CUDA and TensorRT), we selected CUDA for our GPU-based environment (i.e., Cloud) due to its notable inference benefits[12][13] over TensorRT. Nonetheless, relying on CUDA introduces certain limitations. Framework-specific optimizations, hardware-specific dependencies, and differences in precision or ecosystem support could affect the generalizability of our

[12] https://developer.nvidia.com/blog/end-to-end-ai-for-nvidia-based-pcs-cuda-and-tensorrt-execution-providers-in-onnx-runtime

[13] https://github.com/chaiNNer-org/chaiNNer/discussions/2437

results. For example, TensorRT, while optimized for high-performance inference, may outperform CUDA in scenarios with static model execution graphs. Furthermore, CUDA's dependency on NVIDIA GPUs limits our findings, potentially biasing performance outcomes compared to hardware like AMD or ARM-based systems.

Despite these constraints, there were several reasons for this choice. First, CUDA is faster to initialize as it evaluates only small network building blocks during its exhaustive search, leveraging the cuDNN inference library for granular neural network operations. This makes CUDA more versatile in handling inference tasks with varying sizes, easier to use, and less demanding in terms of setup. In contrast, TensorRT evaluates entire graphs and explores all execution paths, which can take several minutes for large ONNX models. Additionally, TensorRT's allocation of workspace memory for intermediate buffers leads to higher memory usage, and its frequent engine recomputation for varying input sizes—such as in our study involving diverse image sizes—can result in slower performance. Thus, alternative providers like TensorRT were deemed suboptimal for our specific requirements.

Instead of using actual devices, we used Docker containers for simulating the hardware and network configurations of physical Mobile, Edge, and Cloud devices to create real-time deployment scenarios. The Docker simulations allow flexibility by easily configuring the network/hardware settings. Setting up and maintaining actual devices can be expensive and require more experience, especially for simulating different configurations and deployment scenarios. Docker containers provide a cost-effective way to create virtual environments that closely mimic the behavior of real hardware and network bandwidth configurations without the need for additional physical resources. The latest versions of all the tools and packages were employed on the simulated devices in the experimental setup (Section 4). The generalization factor can be improved by replicating the experiment on different hardware and network configurations of the devices. In other words, real-world deployment considerations, such as network variability, security implications, or dynamic Edge environments, should be considered for generalization.

In our study, we computed the inference accuracy performance independently on multiple representative deployment tiers (i.e., Mobile, Edge, Cloud) for four different types of accuracy-sensitive operators (i.e., Identity, Quantized, Early Exiting, and Quantized Early Exiting) to provide valuable insights into the model's generalizability across different hardware targets, specifically CPUs and GPUs. The ONNX models are designed to be hardware-agnostic and can be deployed on various hardware devices without significant modifications. This allows the models to achieve consistent accuracy performance across different deployment environments, as long as the hardware supports the necessary operations and computational capabilities.

We applied the operators to the subjects in ONNX format and performed the inference of the transformed models using the ONNX Runtime Engine due to optimized deployment performance benefits, as suggested by previous studies [86,38]. The feasibility of scripting black-box transformations using Python ONNX APIs was another reason for considering subject models in ONNX format, instead of other formats (like Pytorch and Tensorflow).

8 Conclusions and Future Work

Deploying black-box models (DNNs) efficiently in an Edge AI setting introduces unique challenges for MLOps Engineers and software practitioners. The black-box models require specific considerations for optimization in resource-constrained and network-constrained deployment scenarios. This paper aims to be an important stepping stone in the field of MLOps, in particular for the deployment of black-box models, to evaluate the benefits and trade-offs of Edge AI deployment strategies involving mappings of <operators, tiers>, by evaluating their performance in terms of quantitative metrics like latency and accuracy. While previous works focused on exploring and addressing individual operators (i.e., Partitioned, Early Exit, Quantization), our study has systematically compared the individual operators and their unexplored combinations in an Edge AI Environment using empirical data of four major CV subjects for testing the various deployment strategies.

The MLOps Engineers could prefer Mobile-Edge distributed inference when faster latency is a concern in deployment scenarios where the mobile tier has strict resource (CPU/RAM) requirements. For models with smaller input data size requirements, their deployment at the Cloud tier with limited network bandwidth capacity can also be a better alternative than Model Partitioned across Mobile, Edge, and Cloud tiers and Mobile/Edge deployment. For models with large input data size requirements, Edge deployment can be a priority over Model Partitioned and Mobile/Cloud deployment in scenarios where the Edge has higher computational/network capabilities than the Mobile/Cloud. Among the studied operators, the Edge deployment of the Quantized Early Exit operator could be the preferred choice over the Edge deployment of the Early Exit/Quantized operator and Mobile-Edge deployment of the Partition operator when faster latency is a requirement at medium accuracy loss. In contrast, for MLOps Engineers having requirements of the minimal accuracy loss w.r.t the original model, the Edge deployment of the Quantized operator could be the preferred choice at the benefit of faster latency over the Edge/Mobile-Edge deployment of the Early Exit/Partition operator and the cost of slower latency over the Edge deployment of Quantized Early Exit operator. Deploying Non-Partitioned models with small input data size (i.e., FCN) is ideal for Cloud deployment even in bandwidth-constrained settings ($\leq$10 Mbps). Whereas, deploying the Non-Partitioned models with large input data size (ResNe(x)t, DUC) is suitable for Cloud deployment with moderate bandwidth availability ($\geq$50 Mbps). For models with higher intermediate data sizes (i.e., FCN, DUC), the Partition-based strategies need higher bandwidths ($\geq$50 Mbps) for latency convergence. For Non-Partitioned models with large input data sizes (ResNe(x)t), the Mobile and Edge deployment latencies converge at 50 Mbps. In general, the Cloud tier outperforms the Edge and Mobile tier for the Non-Partitioning operators when MEC bandwidth is at least 50 Mbps, but remains suboptimal under lower bandwidth conditions. Additionally, Mobile-Edge Partitioning-based strategies latency performance consistently exceeds Mobile-Cloud and Edge-Cloud alternatives.

Our study focuses on the impact of network bandwidth variations across Mobile-Edge and Edge-Cloud environments, as this is a critical factor for latency performance evaluation. However, system-level factors such as CPU/memory usage, network throughput, and energy consumption were not included in our current analysis. These factors may also influence overall latency performance and merit further investigation. While we scoped this study to specifically analyze the effects of network bandwidth, future work could incorporate these additional variables to provide a more holistic evaluation of system performance under varying resource constraints. The provided empirical results for these operators on the four Image Classification and Segmentation subjects give valuable insights into the speed and performance of operators across and within tiers under various deployment scenarios, which could inspire work related to the automation of these operators for future studies in the field of MLOps. Additionally, the empirical approach employed in our study, and the empirical results obtained, can be used as a foundation for developing recommendation systems for Edge AI operators.

9 Compliance with Ethical Standards

Conflict of Interest: All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
Funding: This study was funded by NSERC.
Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.
Informed Consent: Not applicable.

10 Data Availability Statement

The models generated and analyzed during the current study are available at the following link:
https://github.com/SAILResearch/wip-24-jaskirat-black-box-edge-operators.git.

References

1. Alqarni, M., Cherif, A., and Alkayyal, E. Odm-bcsa: An offloading decision-making framework based on binary cuckoo search algorithm for mobile edge computing. *Computer Networks 226* (2023), 109647.
2. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data 8* (2021), 1–74.

3. Andrés Ramiro, C., Fiandrino, C., Blanco Pizarro, A., Jiménez Mateo, P., Ludant, N., and Widmer, J. openleon: An end-to-end emulator from the edge data center to the mobile users. In *Proceedings of the 12th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization* (2018), pp. 19–27.
4. Assine, J. S., Valle, E., et al. Single-training collaborative object detectors adaptive to bandwidth and computation. *arXiv preprint arXiv:2105.00591* (2021).
5. Banner, R., Nahshan, Y., and Soudry, D. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems 32* (2019).
6. Basili, V. R. Applying the goal/question/metric paradigm in the experience factory. *Software quality assurance and measurement: A worldwide perspective 7*, 4 (1993), 21–44.
7. Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13169–13178.
8. Chen, T., Moreau, T., Jiang, Z., Zheng, L., Yan, E., Shen, H., Cowan, M., Wang, L., Hu, Y., Ceze, L., et al. {TVM}: An automated {End-to-End} optimizing compiler for deep learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)* (2018), pp. 578–594.
9. Chiang, C.-H., Liu, P., Wang, D.-W., Hong, D.-Y., and Wu, J.-J. Optimal branch location for cost-effective inference on branchynet. In *2021 IEEE International Conference on Big Data (Big Data)* (2021), IEEE, pp. 5071–5080.
10. Choi, H., and Bajić, I. V. Deep feature compression for collaborative object detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), IEEE, pp. 3743–3747.
11. Choi, H., Cohen, R. A., and Bajić, I. V. Back-and-forth prediction for deep tensor compression. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 4467–4471.
12. Choukroun, Y., Kravchik, E., Yang, F., and Kisilev, P. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019), IEEE, pp. 3009–3018.
13. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 3 (1993), 494.
14. Cohen, R. A., Choi, H., and Bajić, I. V. Lightweight compression of neural network feature tensors for collaborative intelligence. In *2020 IEEE International Conference on Multimedia and Expo (ICME)* (2020), IEEE, pp. 1–6.
15. Conover, W. J., and Iman, R. L. On multiple-comparisons procedures. *Los Alamos Sci. Lab. Tech. Rep. LA-7677-MS 1* (1979), 14.
16. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
17. Cui, H., Tang, Z., Lou, J., Jia, W., and Zhao, W. Latency-aware container scheduling in edge cluster upgrades: A deep reinforcement learning approach. *IEEE Transactions on Services Computing* (2024).
18. Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., and Zomaya, A. Y. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal 7*, 8 (2020), 7457–7469.
19. Dimolitsas, I., Spatharakis, D., Dechouniotis, D., Zafeiropoulos, A., and Papavassiliou, S. Multi-application hierarchical autoscaling for kubernetes edge clusters. In *2023 IEEE International Conference on Smart Computing (SMARTCOMP)* (2023), IEEE, pp. 291–296.
20. Dong, C., Hu, S., Chen, X., and Wen, W. Joint optimization with dnn partitioning and resource allocation in mobile edge computing. *IEEE Transactions on Network and Service Management 18*, 4 (2021), 3973–3986.
21. Duan, Y., and Wu, J. Joint optimization of dnn partition and scheduling for mobile cloud computing. In *Proceedings of the 50th International Conference on Parallel Processing* (2021), pp. 1–10.
22. Dwivedi, A. K., Mallawaarachchi, I., and Alvarado, L. A. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine 36*, 14 (2017), 2187–2205.
23. Elbayad, M., Gu, J., Grave, E., and Auli, M. Depth-adaptive transformer. *arXiv preprint arXiv:1910.10073* (2019).
24. Eshratifar, A. E., Abrishami, M. S., and Pedram, M. Jointdnn: An efficient training and inference engine for intelligent mobile cloud computing services. *IEEE Transactions on Mobile Computing 20*, 2 (2019), 565–576.
25. Eshratifar, A. E., Esmaili, A., and Pedram, M. Bottlenet: A deep learning architecture for intelligent mobile cloud computing services. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* (2019), IEEE, pp. 1–6.
26. Fan, A., Stock, P., Graham, B., Grave, E., Gribonval, R., Jegou, H., and Joulin, A. Training with quantization noise for extreme model compression. *arXiv preprint arXiv:2004.07320* (2020).
27. Fang, J., Shafiee, A., Abdel-Aziz, H., Thorsley, D., Georgiadis, G., and Hassoun, J. H. Post-training piecewise linear quantization for deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, 2020, Proceedings, Part II 16* (2020), Springer, pp. 69–86.
28. Fiandrino, C., Pizarro, A. B., Mateo, P. J., Ramiro, C. A., Ludant, N., and Widmer, J. openleon: An end-to-end emulation platform from the edge data center to the mobile user. *Computer Communications 148* (2019), 17–26.
29. Garg, S., Jain, A., Lou, J., and Nahmias, M. Confounding tradeoffs for neural network quantization. *arXiv preprint arXiv:2102.06366* (2021).
30. Garg, S., Lou, J., Jain, A., Guo, Z., Shastri, B. J., and Nahmias, M. Dynamic precision analog computing for neural networks. *IEEE Journal of Selected Topics in Quantum Electronics 29*, 2: Optical Computing (2022), 1–12.
31. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022, pp. 291–326.
32. Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

33. Görmez, A., Dasari, V. R., and Koyuncu, E. E 2 cm: Early exit via class means for efficient supervised and unsupervised learning. In *2022 International Joint Conference on Neural Networks (IJCNN)* (2022), IEEE, pp. 1–8.
34. Görmez, A., and Koyuncu, E. Class based thresholding in early exit semantic segmentation networks. *arXiv preprint arXiv:2210.15621* (2022).
35. Gray, R. M., and Neuhoff, D. L. Quantization. *IEEE transactions on Information Theory 44*, 6 (1998), 2325–2383.
36. Greengard, S. Ai on edge. *Commun. ACM 63*, 9 (2020), 18–20.
37. Guadagnoli, E., and Velicer, W. F. Relation of sample size to the stability of component patterns. *Psychological Bulletin 103*, 2 (1988), 265.
38. Hampau, R. M., Kaptein, M., Van Emden, R., Rost, T., and Malavolta, I. An empirical study on the performance and energy consumption of ai containerization strategies for computer-vision tasks on the edge. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022* (2022), pp. 50–59.
39. Hawks, B., Duarte, J., Fraser, N. J., Pappalardo, A., Tran, N., and Umuroglu, Y. Ps and qs: Quantization-aware pruning for efficient low latency neural network inference. *Frontiers in Artificial Intelligence 4* (2021), 676564.
40. He, X., and Cheng, J. Learning compression from limited unlabeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 752–769.
41. Hess, M. R., and Kromrey, J. D. Robust confidence intervals for effect sizes: A comparative study of cohen'sd and cliff's delta under non-normality and heterogeneous variances. In *Annual Meeting of the American Educational Research Association* (2004), vol. 1, Citeseer.
42. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* (1979), 65–70.
43. Hu, C., Bao, W., Wang, D., and Liu, F. Dynamic adaptive dnn surgery for inference acceleration on the edge. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications* (2019), IEEE, pp. 1423–1431.
44. Hu, D., and Krishnamachari, B. Fast and accurate streaming cnn inference via communication compression on the edge. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)* (2020), IEEE, pp. 157–163.
45. Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., and Soudry, D. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518* (2020).
46. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2704–2713.
47. Jankowski, M., Gündüz, D., and Mikolajczyk, K. Joint device-edge inference over wireless links with pruning. In *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)* (2020), IEEE, pp. 1–5.
48. Jeong, H.-J., Jeong, I., Lee, H.-J., and Moon, S.-M. Computation offloading for machine learning web apps in the edge server environment. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)* (2018), IEEE, pp. 1492–1499.
49. Jiang, Z. M., and Hassan, A. E. A survey on load testing of large-scale software systems. *IEEE Transactions on Software Engineering 41*, 11 (2015), 1091–1118.
50. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., and Tang, L. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News 45*, 1 (2017), 615–629.
51. Krishnamoorthi, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342* (2018).
52. Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 25* (2012).
53. Künas, C. A. Optimizing machine learning models training in the cloud. Master's thesis, Universidade Federal do Rio Grande do Sul, 2023.
54. Laskaridis, S., Venieris, S. I., Almeida, M., Leontiadis, I., and Lane, N. D. Spinn: synergistic progressive inference of neural networks over device and cloud. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (2020), pp. 1–15.
55. LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521*, 7553 (2015), 436–444.
56. Lee, J. C., Kim, Y., Moon, S., and Ko, J. H. A splittable dnn-based object detector for edge-cloud collaborative real-time video inference. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2021), IEEE, pp. 1–8.
57. Lee, J. H., Ha, S., Choi, S., Lee, W.-J., and Lee, S. Quantization for rapid deployment of deep neural networks. *arXiv preprint arXiv:1810.05488* (2018).
58. Li, E., Zhou, Z., and Chen, X. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings of the 2018 Workshop on Mobile Edge Communications* (2018), pp. 31–36.
59. Li, G., Liu, L., Wang, X., Dong, X., Zhao, P., and Feng, X. Auto-tuning neural network quantization framework for collaborative inference between the cloud and edge. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Proceedings, Part I 27* (2018), Springer, pp. 402–411.
60. Li, H., Zhang, H., Qi, X., Yang, R., and Huang, G. Improved techniques for training adaptive deep networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1891–1900.
61. Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426* (2021).
62. Liang, H., Sang, Q., Hu, C., Cheng, D., Zhou, X., Wang, D., Bao, W., and Wang, Y. Dnn surgery: Accelerating dnn inference on the edge through layer partitioning. *IEEE Transactions on Cloud Computing 11*, 3 (2023), 3111–3125.
63. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, 2014, Proceedings, Part V 13* (2014), Springer, pp. 740–755.

64. Liu, D., Kong, H., Luo, X., Liu, W., and Subramaniam, R. Bringing ai to edge: From deep learning's perspective. *Neurocomputing 485* (2022), 297–320.
65. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Deng, H., and Ju, Q. Fastbert: a self-distilling bert with adaptive inference time. *arXiv preprint arXiv:2004.02178* (2020).
66. Lo, C., Su, Y.-Y., Lee, C.-Y., and Chang, S.-C. A dynamic deep neural network design for efficient workload allocation in edge computing. In *2017 IEEE International Conference on Computer Design (ICCD)* (2017), IEEE, pp. 273–280.
67. Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation, 2015.
68. Lu, D., and Weng, Q. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing 28*, 5 (2007), 823–870.
69. Luo, D., Yu, T., Wu, Y., Wu, H., Wang, T., and Zhang, W. Split: Qos-aware dnn inference on shared gpu via evenly-sized model splitting. In *Proceedings of the 52nd International Conference on Parallel Processing* (2023), pp. 605–614.
70. Matsubara, Y., Baidya, S., Callegaro, D., Levorato, M., and Singh, S. Distilled split deep neural networks for edge-assisted real-time systems. In *Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges* (2019), pp. 21–26.
71. Matsubara, Y., Callegaro, D., Singh, S., Levorato, M., and Restuccia, F. Bottlefit: Learning compressed representations in deep neural networks for effective and efficient split computing. In *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (2022), IEEE, pp. 337–346.
72. Matsubara, Y., and Levorato, M. Split computing for complex object detectors: Challenges and preliminary results. In *Proceedings of the 4th International Workshop on Embedded and Mobile Deep Learning* (2020), pp. 7–12.
73. Matsubara, Y., and Levorato, M. Neural compression and filtering for edge-assisted real-time object detection in challenged networks. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 2272–2279.
74. Matsubara, Y., Levorato, M., and Restuccia, F. Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys 55*, 5 (2022), 1–30.
75. Matsubara, Y., Yang, R., Levorato, M., and Mandt, S. Sc2 benchmark: Supervised compression for split computing. *arXiv e-prints* (2022), arXiv–2203.
76. Matsubara, Y., Yang, R., Levorato, M., and Mandt, S. Supervised compression for resource-constrained edge computing systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022), pp. 2685–2695.
77. Meller, E., Finkelstein, A., Almog, U., and Grobman, M. Same, same but different: Recovering neural network quantization error through weight factorization. In *International Conference on Machine Learning* (2019), PMLR, pp. 4486–4495.
78. Merkel, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J. 2014*, 239 (mar 2014).
79. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 7 (2021), 3523–3542.
80. Mohammed, T., Joe-Wong, C., Babbar, R., and Di Francesco, M. Distributed inference acceleration with adaptive dnn partitioning and offloading. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications* (2020), IEEE, pp. 854–863.
81. Murshed, M. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., and Hussain, F. Machine learning at the network edge: A survey. *ACM Computing Surveys 54*, 8 (2021), 1–37.
82. Na, J., Zhang, H., Lian, J., and Zhang, B. Genetic algorithm-based online-partitioning branchynet for accelerating edge inference. *Sensors 23*, 3 (2023), 1500.
83. Nagel, M., Baalen, M. v., Blankevoort, T., and Welling, M. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1325–1334.
84. Nan, Y., Jiang, S., and Li, M. Large-scale video analytics with cloud–edge collaborative continuous learning. *ACM Transactions on Sensor Networks 20*, 1 (2023), 1–23.
85. Neshatpour, K., Homayoun, H., and Sasan, A. Icnn: The iterative convolutional neural network. *ACM Transactions on Embedded Computing Systems (TECS) 18*, 6 (2019), 1–27.
86. Openja, M., Nikanjam, A., Yahmed, A. H., Khomh, F., Ming, Z., et al. An empirical study of challenges in converting deep learning models. *arXiv preprint arXiv:2206.14322* (2022).
87. Ostertagova, E., Ostertag, O., and Kováč, J. Methodology and application of the kruskal-wallis test. In *Applied Mechanics and Materials* (2014), vol. 611, pp. 115–120.
88. Pagliari, D. J., Chiaro, R., Macii, E., and Poncino, M. Crime: Input-dependent collaborative inference for recurrent neural networks. *IEEE Transactions on Computers 70*, 10 (2020), 1626–1639.
89. Pahl, C. Containerization and the paas cloud. *IEEE Cloud Computing 2*, 3 (2015), 24–31.
90. Phuong, M., and Lampert, C. H. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1355–1364.
91. Pomponi, J., Scardapane, S., and Uncini, A. A probabilistic re-intepretation of confidence scores in multi-exit models. *Entropy 24*, 1 (2021), 1.
92. Portabales, A. R., and Nores, M. L. Dockemu: Extension of a scalable network simulation framework based on docker and ns3 to cover iot scenarios. In *SIMULTECH* (2018), pp. 175–182.
93. Portabales, A. R., and Nores, M. L. Dockemu: An iot simulation framework based on linux containers and the ns-3 network simulator—application to coap iot scenarios. In *Simulation and Modeling Methodologies, Technologies and Applications: 8th International Conference, SIMULTECH 2018, 2018, Revised Selected Papers* (2020), Springer, pp. 54–82.

94. Preuveneers, D., and Joosen, W. Towards multi-party policy-based access control in federations of cloud and edge microservices. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (2019), IEEE, pp. 29–38.
95. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252.
96. Sakr, C., Dai, S., Venkatesan, R., Zimmer, B., Dally, W., and Khailany, B. Optimal clipping and magnitude-aware differentiation for improved quantization-aware training. In *International Conference on Machine Learning* (2022), PMLR, pp. 19123–19138.
97. Sbai, M., Saputra, M. R. U., Trigoni, N., and Markham, A. Cut, distil and encode (cde): Split cloud-edge deep inference. In *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)* (2021), IEEE, pp. 1–9.
98. Scardapane, S., Scarpiniti, M., Baccarelli, E., and Uncini, A. Why should we add early exits to neural networks? *Cognitive Computation 12*, 5 (2020), 954–966.
99. Shao, J., and Zhang, J. Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)* (2020), IEEE, pp. 1–6.
100. Shen, M., Liang, F., Gong, R., Li, Y., Li, C., Lin, C., Yu, F., Yan, J., and Ouyang, W. Once quantization-aware training: High performance extremely low-bit architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5340–5349.
101. Shomron, G., Gabbay, F., Kurzum, S., and Weiser, U. Post-training sparsity-aware quantization. *Advances in Neural Information Processing Systems 34* (2021), 17737–17748.
102. Soldaini, L., and Moschitti, A. The cascade transformer: an application for efficient answer sentence selection. *arXiv preprint arXiv:2005.02534* (2020).
103. Suryavansh, S., Bothra, C., Chiang, M., Peng, C., and Bagchi, S. Tango of edge and cloud execution for reliability. In *Proceedings of the 4th Workshop on Middleware for Edge Clouds & Cloudlets* (2019), pp. 10–15.
104. Tabani, H., Balasubramaniam, A., Arani, E., and Zonooz, B. Challenges and obstacles towards deploying deep learning models on mobile devices. *CoRR abs/2105.02613* (2021).
105. Tailor, S. A., Fernandez-Marques, J., and Lane, N. D. Degree-quant: Quantization-aware training for graph neural networks. *arXiv preprint arXiv:2008.05000* (2020).
106. Teerapittayanon, S., McDanel, B., and Kung, H.-T. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (2016), IEEE, pp. 2464–2469.
107. Teerapittayanon, S., McDanel, B., and Kung, H.-T. Distributed deep neural networks over the cloud, the edge and end devices. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (2017), IEEE, pp. 328–339.
108. Wang, M., Mo, J., Lin, J., Wang, Z., and Du, L. Dynexit: A dynamic early-exit strategy for deep residual networks. In *2019 IEEE International Workshop on Signal Processing Systems (SiPS)* (2019), IEEE, pp. 178–183.
109. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., and Cottrell, G. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), IEEE, pp. 1451–1460.
110. Wang, X., Han, Y., Leung, V. C., Niyato, D., Yan, X., and Chen, X. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials 22*, 2 (2020), 869–904.
111. Wang, Y., Shen, J., Hu, T.-K., Xu, P., Nguyen, T., Baraniuk, R., Wang, Z., and Lin, Y. Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference. *IEEE Journal of Selected Topics in Signal Processing 14*, 4 (2020), 623–633.
112. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., et al. *Experimentation in software engineering*, vol. 236. Springer, 2012.
113. Wołczyk, M., Wójcik, B., Bałazy, K., Podolak, I. T., Tabor, J., Śmieja, M., and Trzcinski, T. Zero time waste: Recycling predictions in early exit neural networks. *Advances in Neural Information Processing Systems 34* (2021), 2516–2528.
114. Wu, H., Judd, P., Zhang, X., Isaev, M., and Micikevicius, P. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602* (2020).
115. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks, 2017.
116. Xin, J., Nogueira, R., Yu, Y., and Lin, J. Early exiting bert for efficient document ranking. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing* (2020), pp. 83–88.
117. Xin, J., Tang, R., Lee, J., Yu, Y., and Lin, J. Deebert: Dynamic early exiting for accelerating bert inference. *arXiv preprint arXiv:2004.12993* (2020).
118. Xing, Q., Xu, M., Li, T., and Guan, Z. Early exit or not: Resource-efficient blind quality enhancement for compressed images. In *European Conference on Computer Vision* (2020), Springer, pp. 275–292.
119. Yang, L., Han, Y., Chen, X., Song, S., Dai, J., and Huang, G. Resolution adaptive networks for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 2369–2378.
120. Yang, X., Chen, D., Qi, Q., Wang, J., Sun, H., Liao, J., and Guo, S. Adaptive dnn surgery for selfish inference acceleration with on-demand edge resource. *arXiv preprint arXiv:2306.12185* (2023).
121. Yao, S., Li, J., Liu, D., Wang, T., Liu, S., Shao, H., and Abdelzaher, T. Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (2020), pp. 476–488.
122. Zagoruyko, S., and Komodakis, N. Wide residual networks, 2017.
123. Zeng, L., Li, E., Zhou, Z., and Chen, X. Boomerang: On-demand cooperative deep neural network inference for edge intelligence on the industrial internet of things. *IEEE Network 33*, 5 (2019), 96–103.

124. Zhang, L. L., Han, S., Wei, J., Zheng, N., Cao, T., Yang, Y., and Liu, Y. Nn-meter: Towards accurate latency prediction of deep-learning model inference on diverse edge devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* (2021), pp. 81–93.
125. Zhang, X., Mounesan, M., and Debroy, S. Effect-dnn: Energy-efficient edge framework for real-time dnn inference. In *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (2023), IEEE, pp. 10–20.
126. Zhao, R., Hu, Y., Dotzel, J., De Sa, C., and Zhang, Z. Improving neural network quantization without retraining using outlier channel splitting. In *International Conference on Machine Learning* (2019), PMLR, pp. 7543–7552.
127. Zhou, L., Wen, H., Teodorescu, R., and Du, D. H. Distributing deep neural networks with containerized partitions at the edge. In *2nd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 19)* (2019).
128. Zhou, W., Xu, C., Ge, T., McAuley, J., Xu, K., and Wei, F. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems 33* (2020), 18330–18341.
129. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., and Zhang, J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE 107*, 8 (2019), 1738–1762.
130. Zhuang, Z., Chen, J., Xu, W., Qi, Q., Guo, S., Wang, J., Lu, L., Yang, H., and Liao, J. Decc: Delay-aware edge-cloud collaboration for accelerating dnn inference. *IEEE Transactions on Emerging Topics in Computing* (2024).

Appendices

Appendix A Graphical Illustrations of Manual Operators

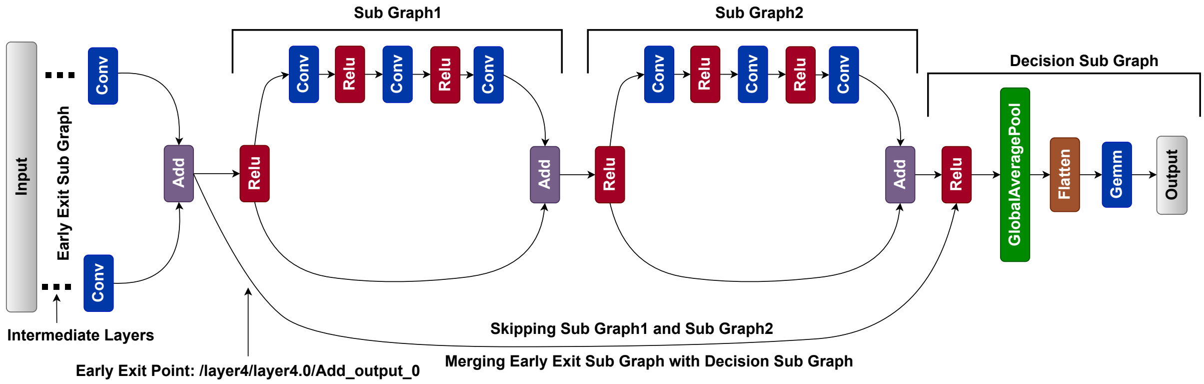


Fig. 15: Graphical Illustration of Early Exiting for ResNet and ResNext

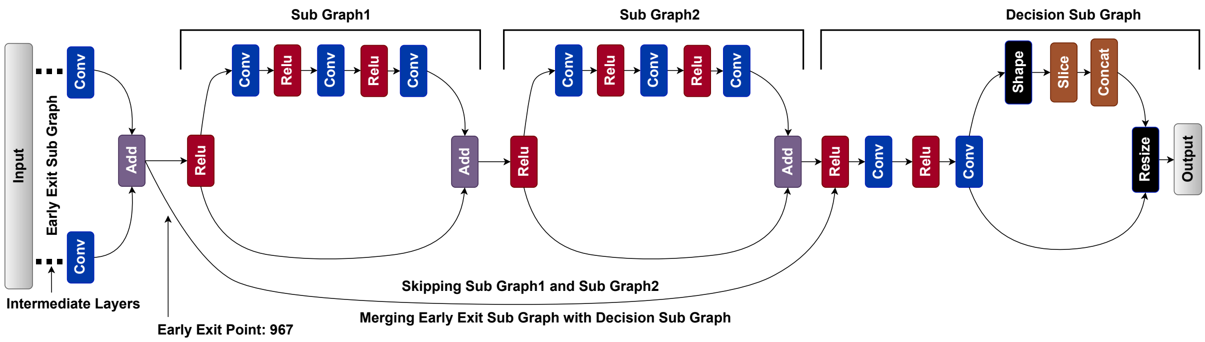


Fig. 16: Graphical Illustration of Early Exiting for FCN

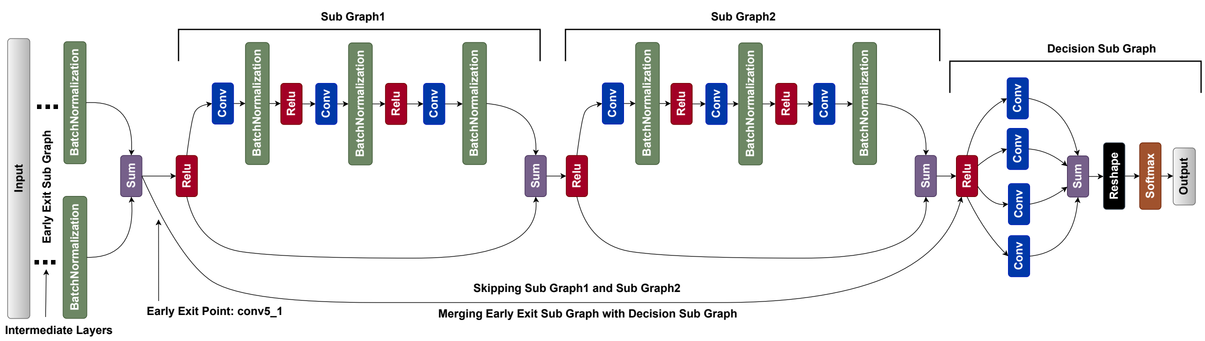


Fig. 17: Graphical Illustration of Early Exiting for DUC

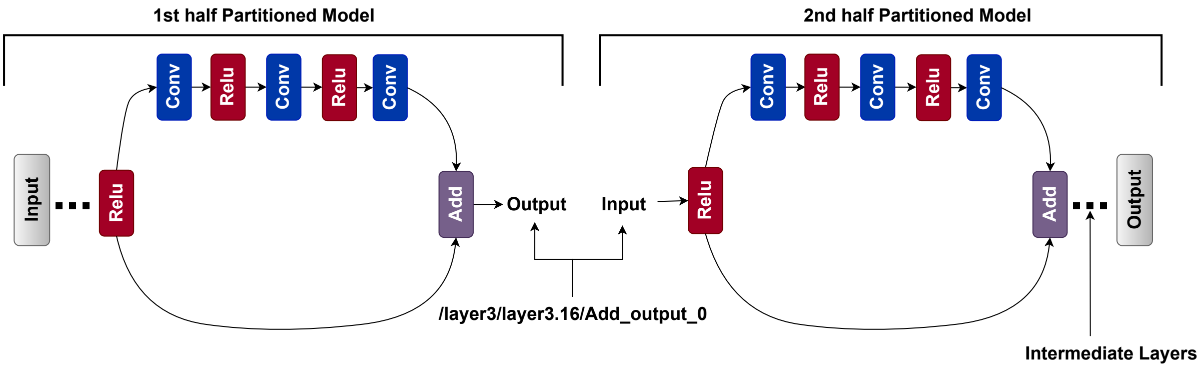


Fig. 18: Graphical Illustration of Model Partitioning for ResNet and ResNext

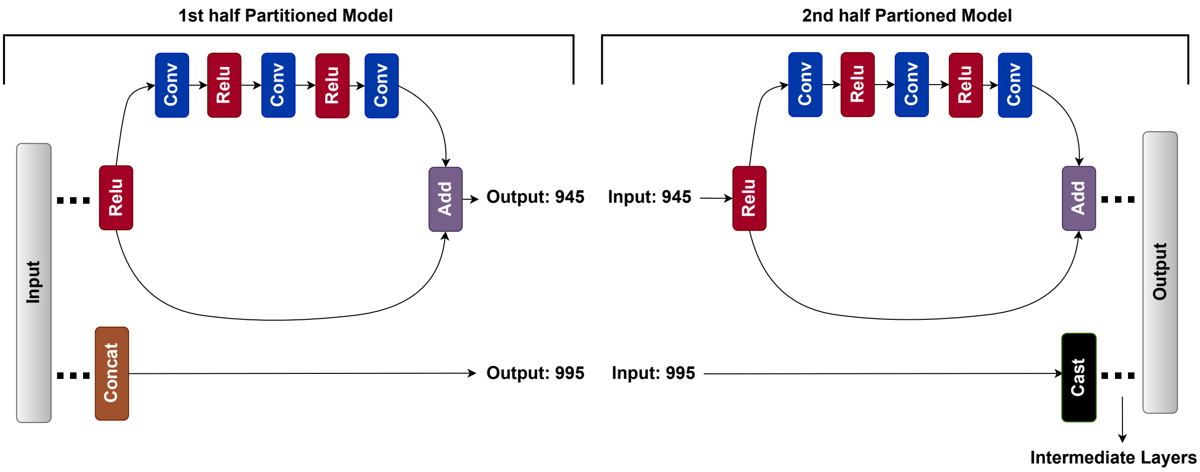


Fig. 19: Graphical Illustration of Model Partitioning for FCN

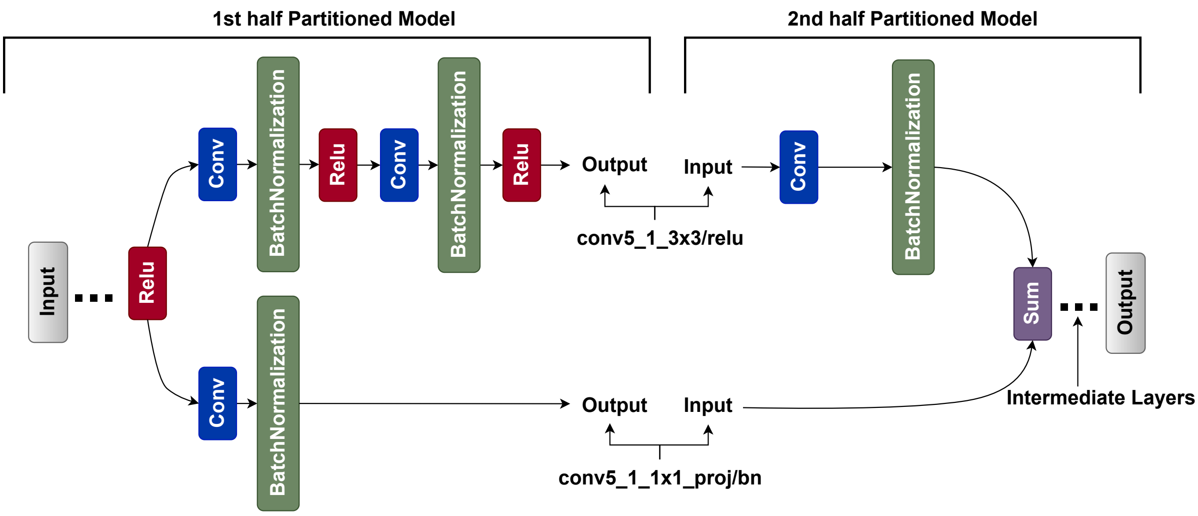


Fig. 20: Graphical Illustration of Model Partitioning for DUC

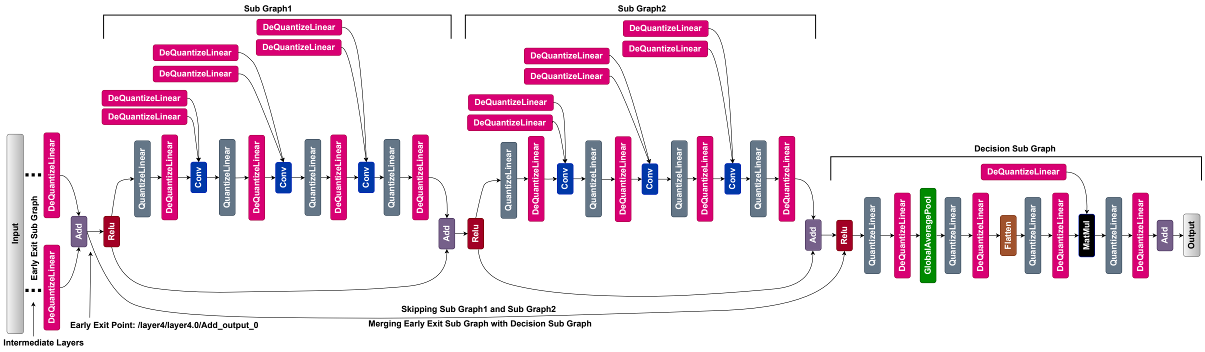


Fig. 21: Graphical Illustration of Quantized Early Exit for ResNet and ResNext

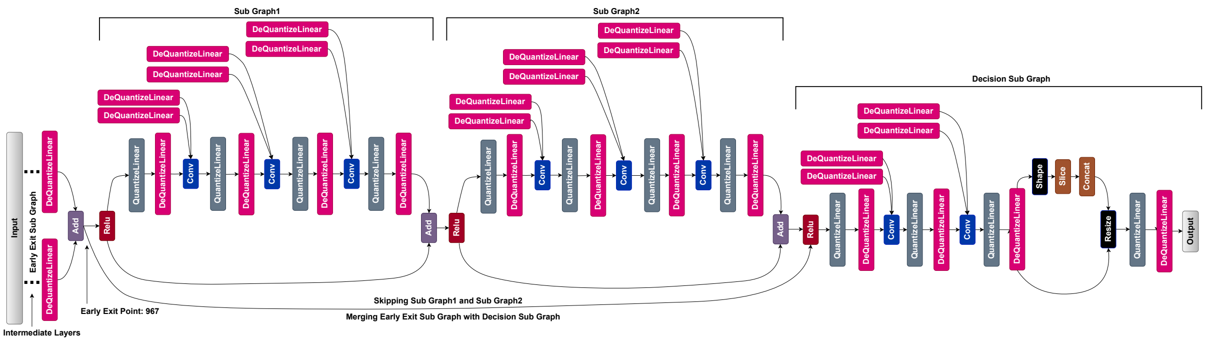


Fig. 22: Graphical Illustration of Quantized Early Exit for FCN

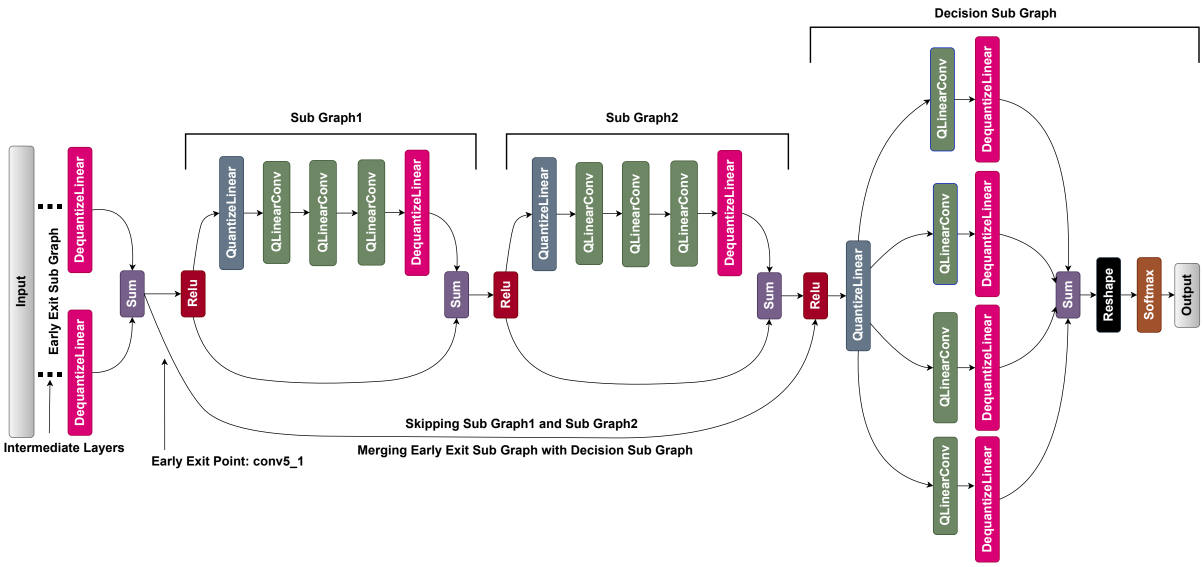


Fig. 23: Graphical Illustration of Quantized Early Exit for DUC

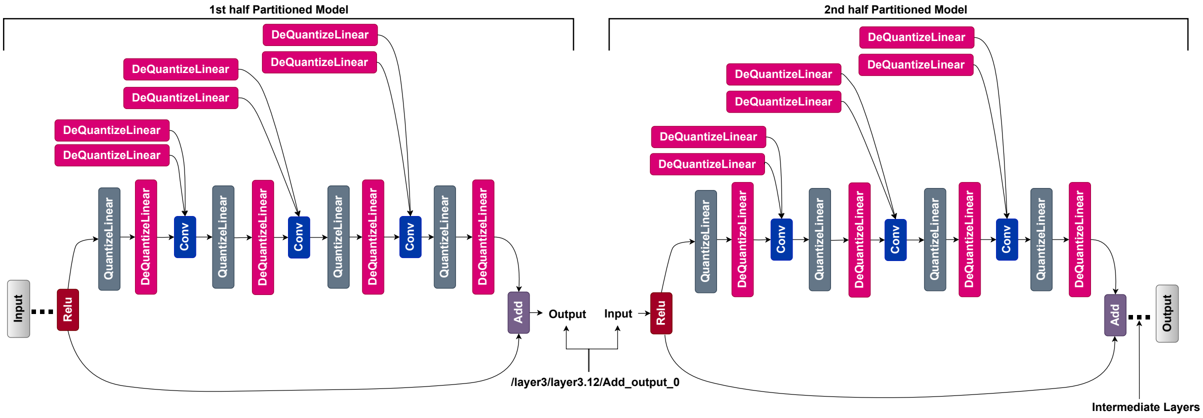


Fig. 24: Graphical Illustration of Quantized Early Exit Partitioning for ResNet and ResNext

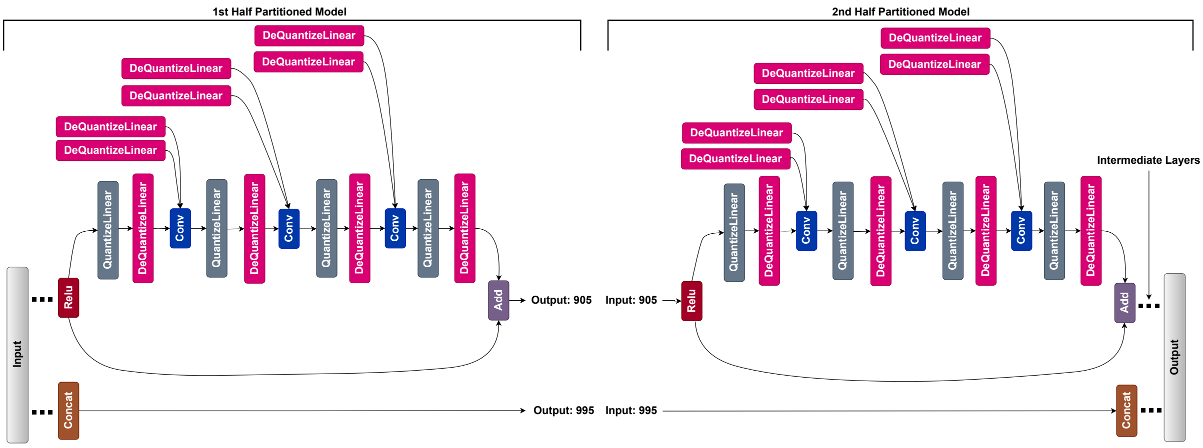


Fig. 25: Graphical Illustration of Quantized Early Exit Partitioning for FCN

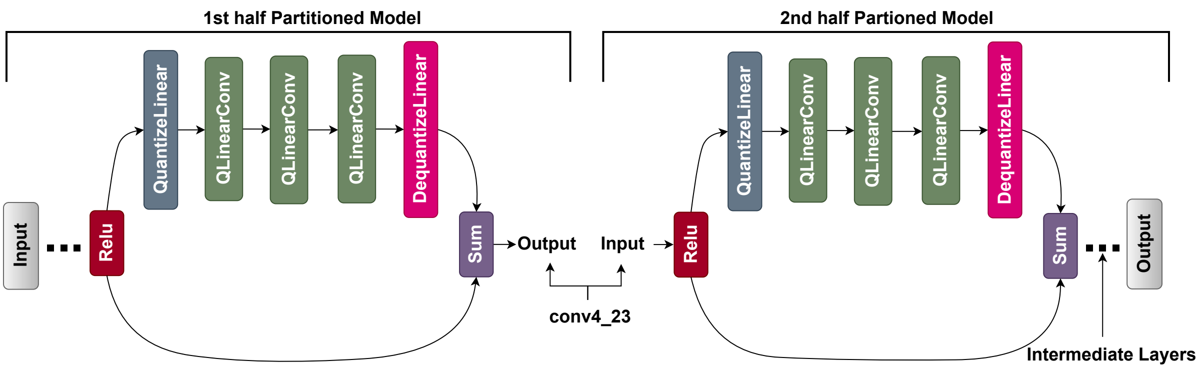


Fig. 26: Graphical Illustration of Quantized Early Exit Partitioning for DUC

Appendix B QQ Plots

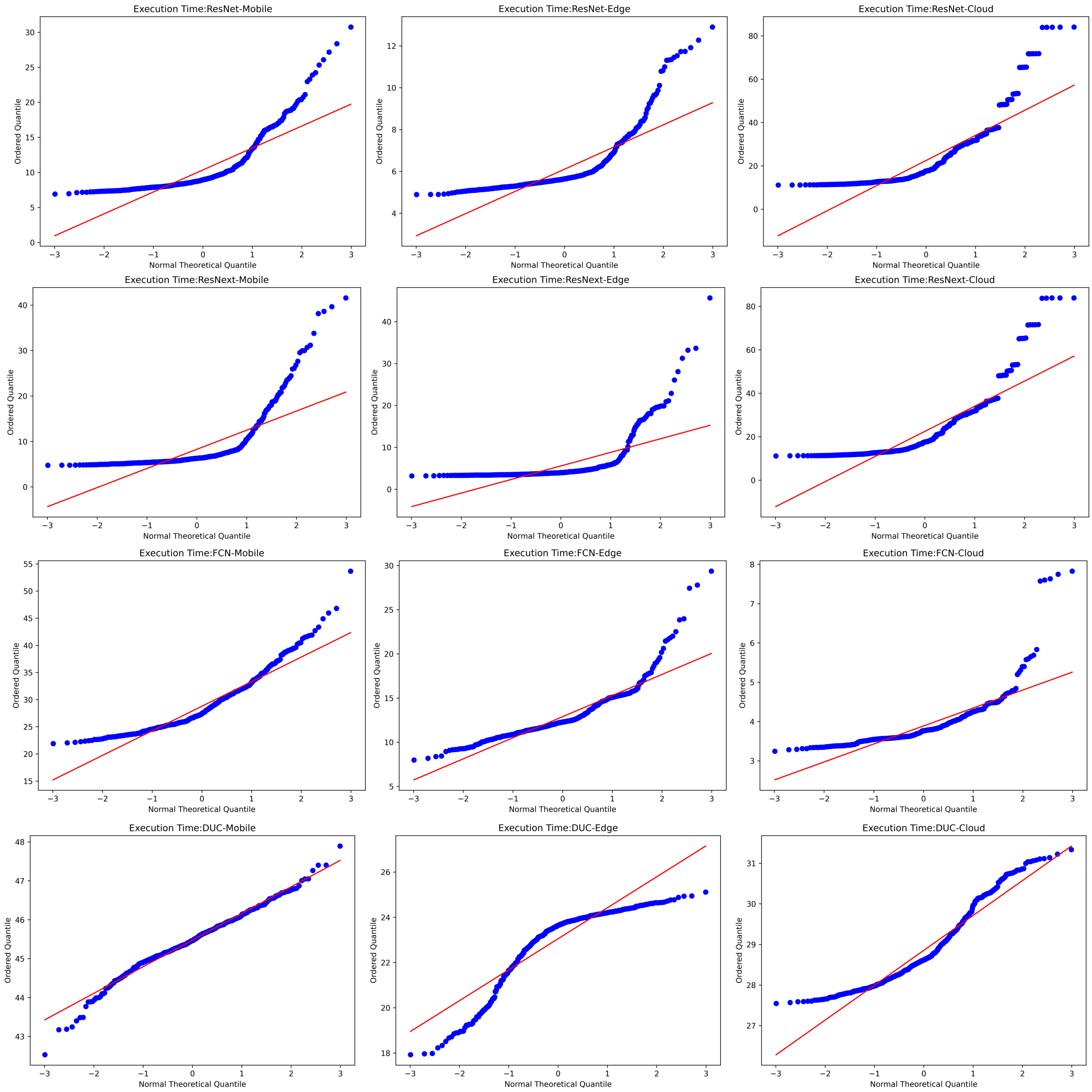


Fig. 27: Graphical Illustration of QQ plots for RQ1 Deployment Strategies

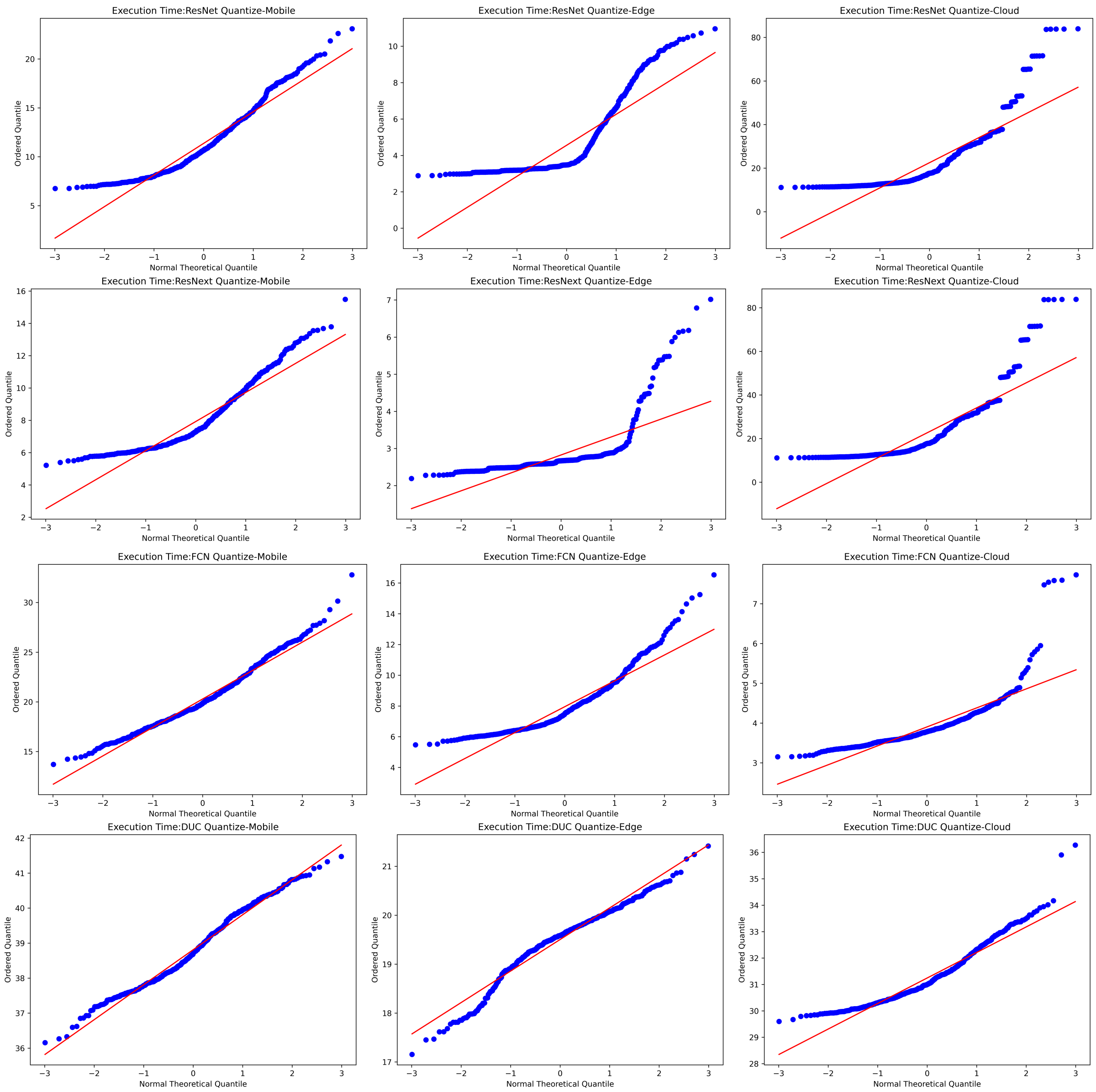


Fig. 28: Graphical Illustration of QQ plots for RQ2 Deployment Strategies

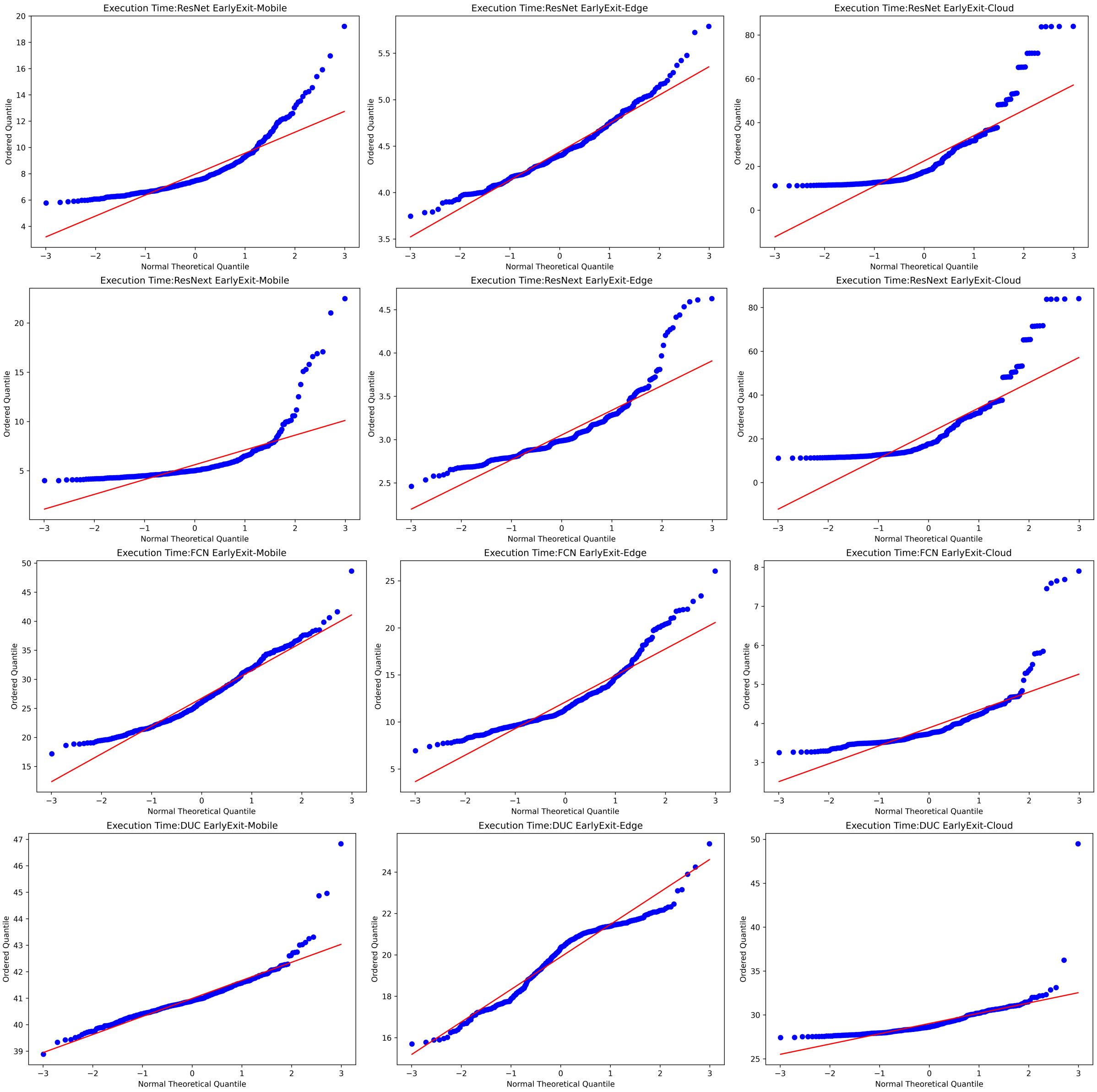
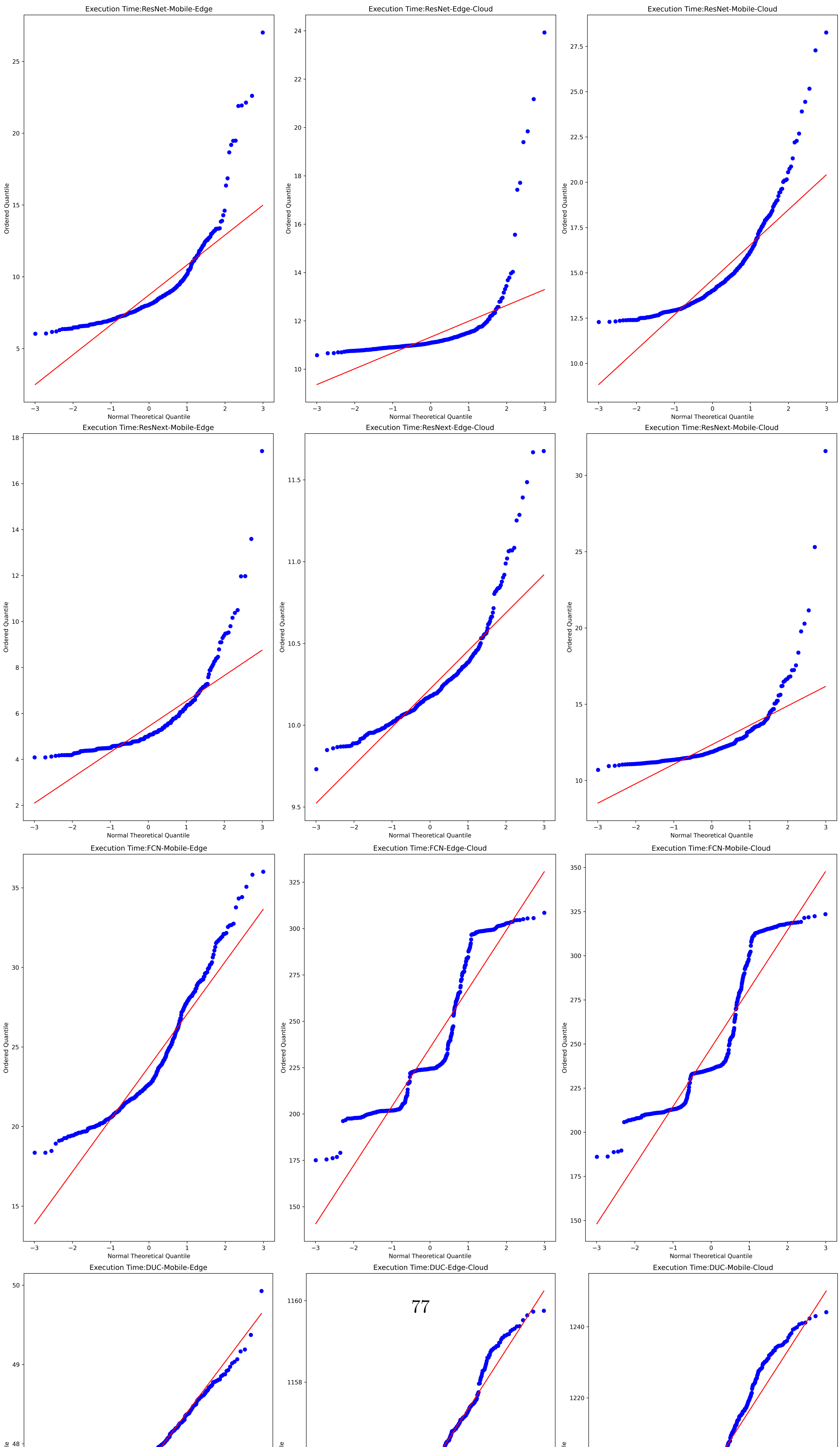


Fig. 29: Graphical Illustration of QQ plots for RQ3 Deployment Strategies


Execution Time:ResNet-Mobile-Edge
Execution Time:ResNet-Edge-Cloud
Execution Time:ResNet-Mobile-Cloud
Execution Time:ResNext-Mobile-Edge
Execution Time:ResNext-Edge-Cloud
Execution Time:ResNext-Mobile-Cloud
Execution Time:FCN-Mobile-Edge
Execution Time:FCN-Edge-Cloud
Execution Time:FCN-Mobile-Cloud
Execution Time:DUC-Mobile-Edge
Execution Time:DUC-Edge-Cloud
Execution Time:DUC-Mobile-Cloud
Ordered Quantile
Normal Theoretical Quantile

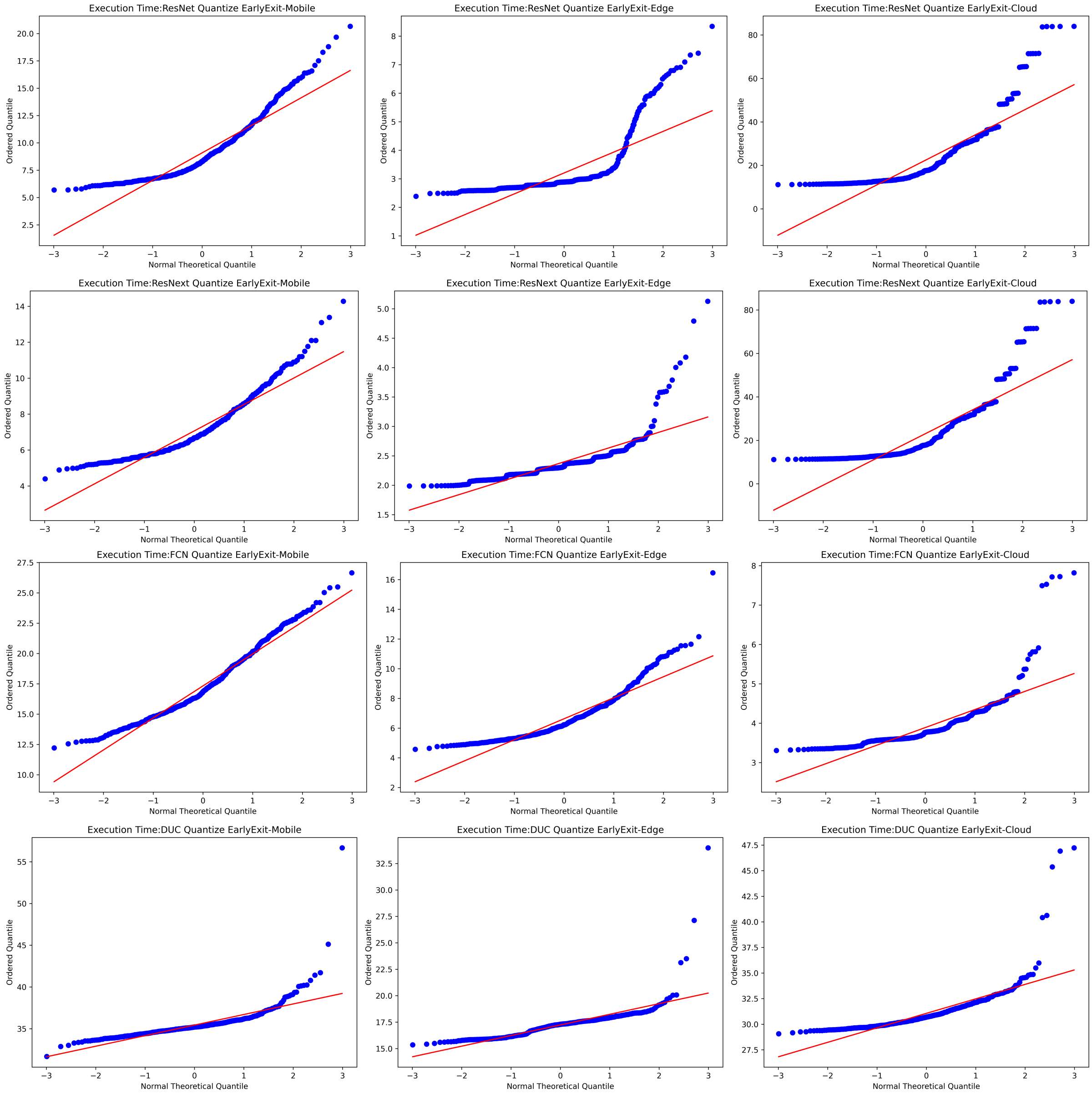
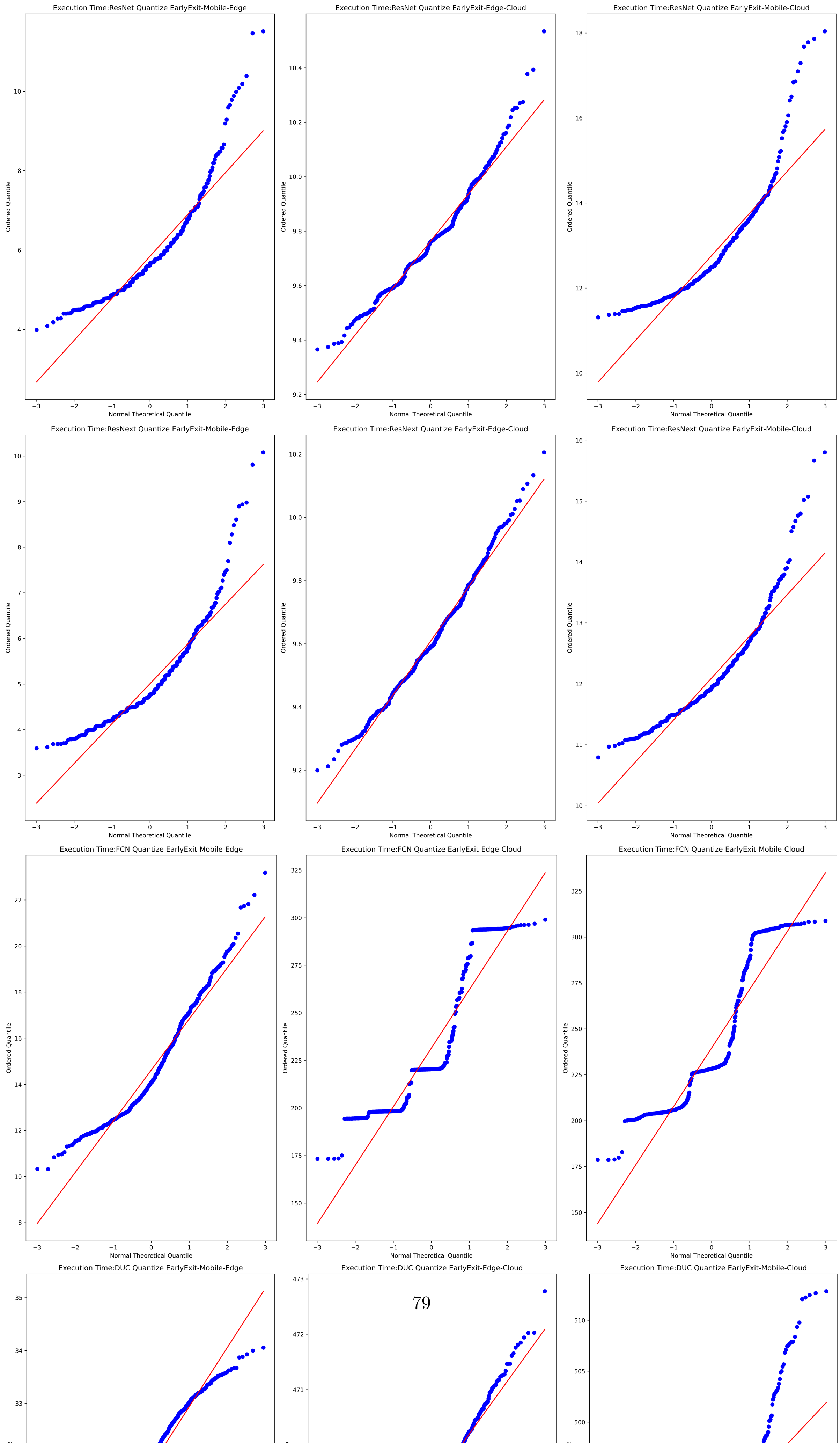


Fig. 31: Graphical Illustration of QQ plots for RQ5 Deployment Strategies


Execution Time:ResNet Quantize EarlyExit-Mobile-Edge
Execution Time:ResNet Quantize EarlyExit-Edge-Cloud
Execution Time:ResNet Quantize EarlyExit-Mobile-Cloud
Execution Time:ResNext Quantize EarlyExit-Mobile-Edge
Execution Time:ResNext Quantize EarlyExit-Edge-Cloud
Execution Time:ResNext Quantize EarlyExit-Mobile-Cloud
Execution Time:FCN Quantize EarlyExit-Mobile-Edge
Execution Time:FCN Quantize EarlyExit-Edge-Cloud
Execution Time:FCN Quantize EarlyExit-Mobile-Cloud
Execution Time:DUC Quantize EarlyExit-Mobile-Edge
Execution Time:DUC Quantize EarlyExit-Edge-Cloud
Execution Time:DUC Quantize EarlyExit-Mobile-Cloud
Ordered Quantile
Normal Theoretical Quantile

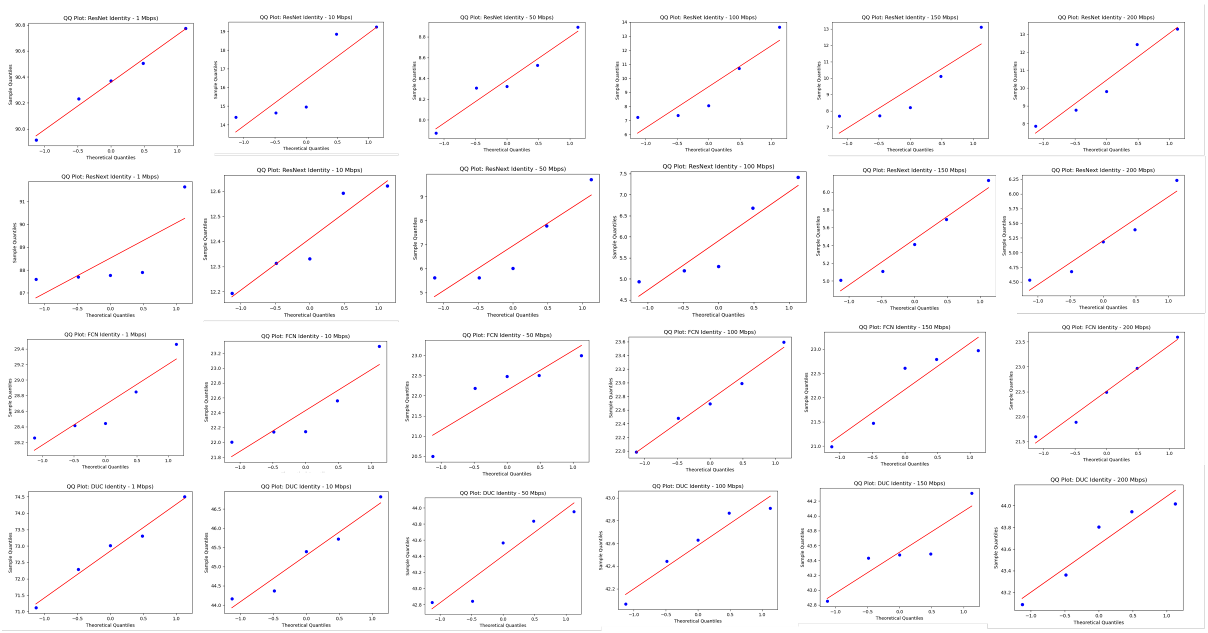


Fig. 33: Graphical Illustration of QQ plots for RQ6 Mobile Identity Deployment Strategies

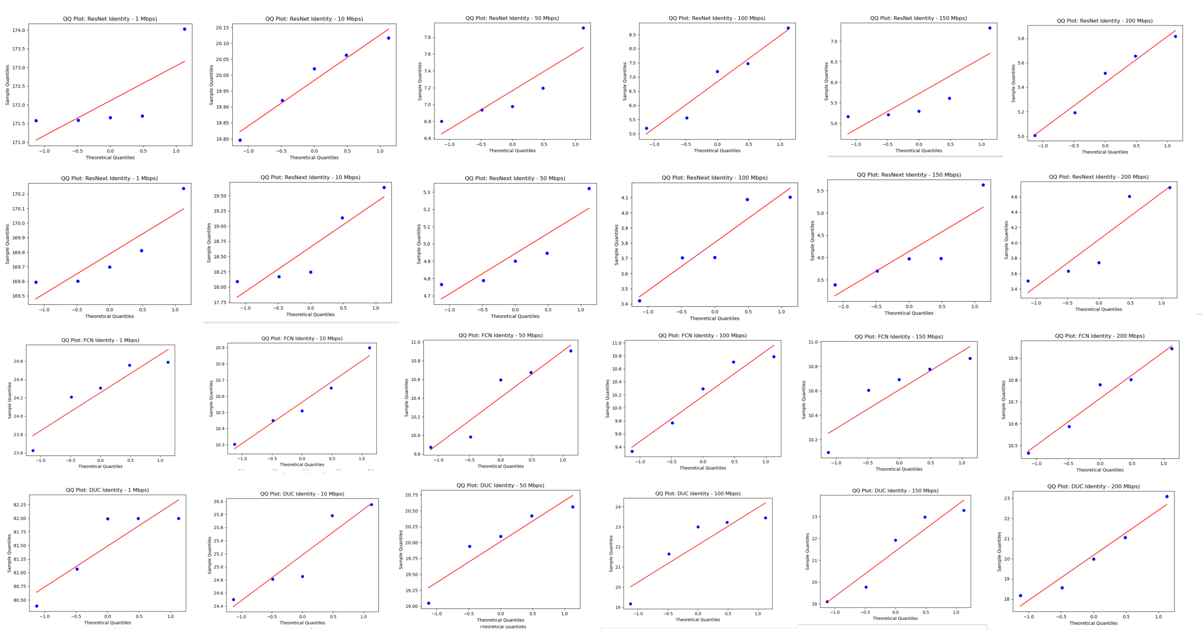


Fig. 34: Graphical Illustration of QQ plots for RQ6 Edge Identity Deployment Strategies

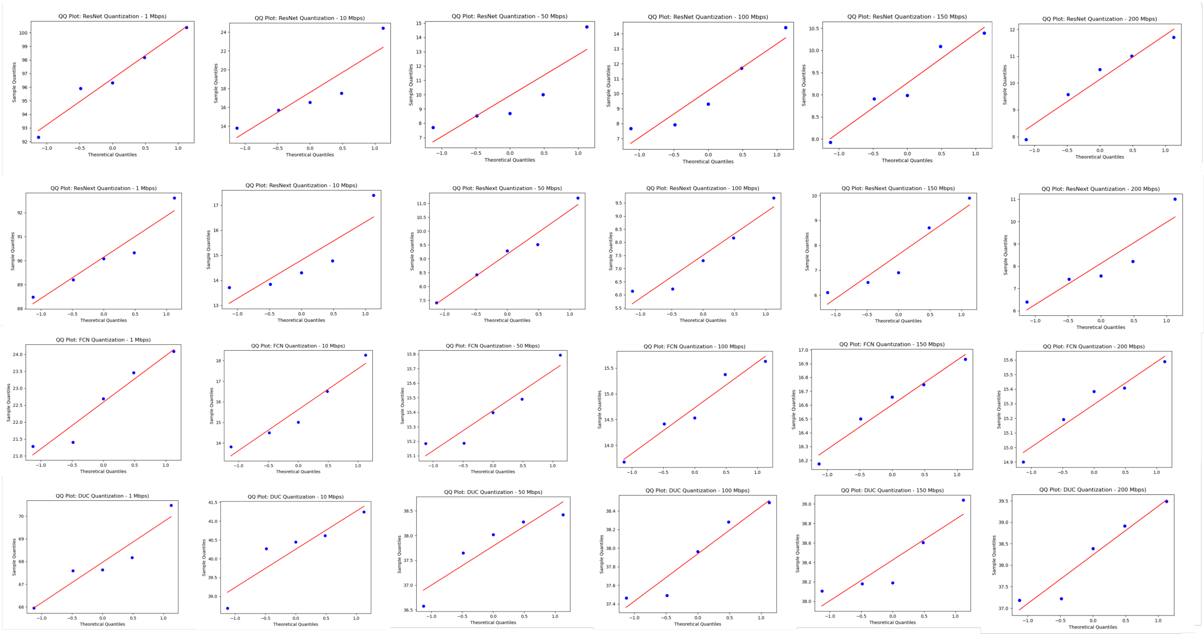


Fig. 35: Graphical Illustration of QQ plots for RQ6 Mobile Quantized Deployment Strategies

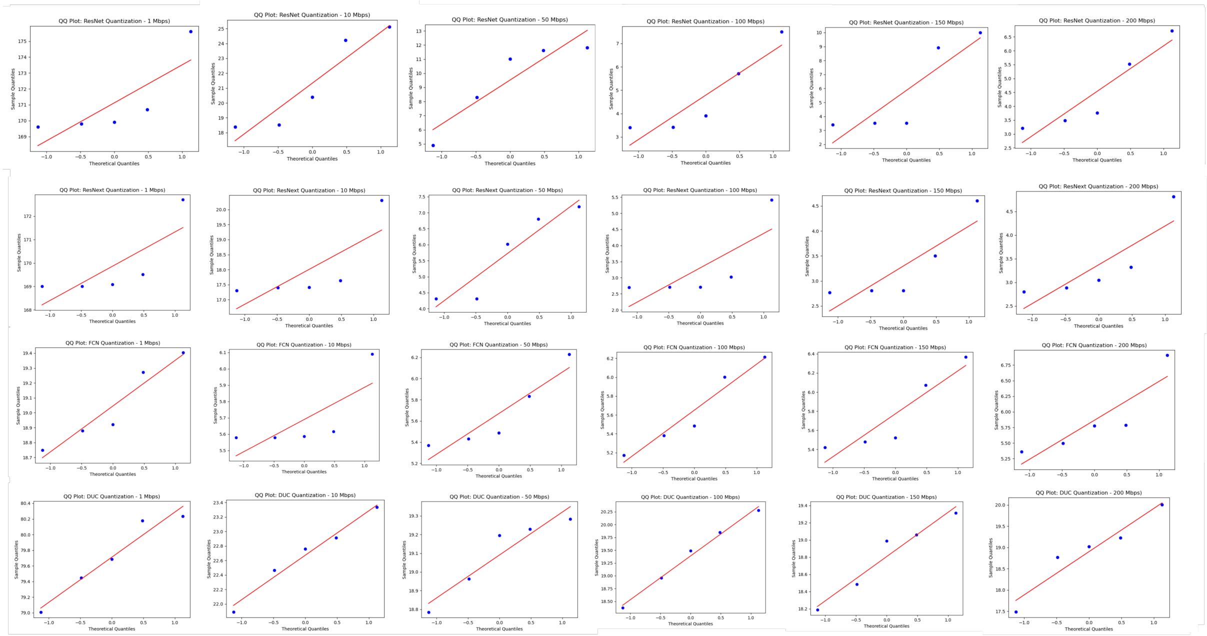


Fig. 36: Graphical Illustration of QQ plots for RQ6 Edge Quantized Deployment Strategies

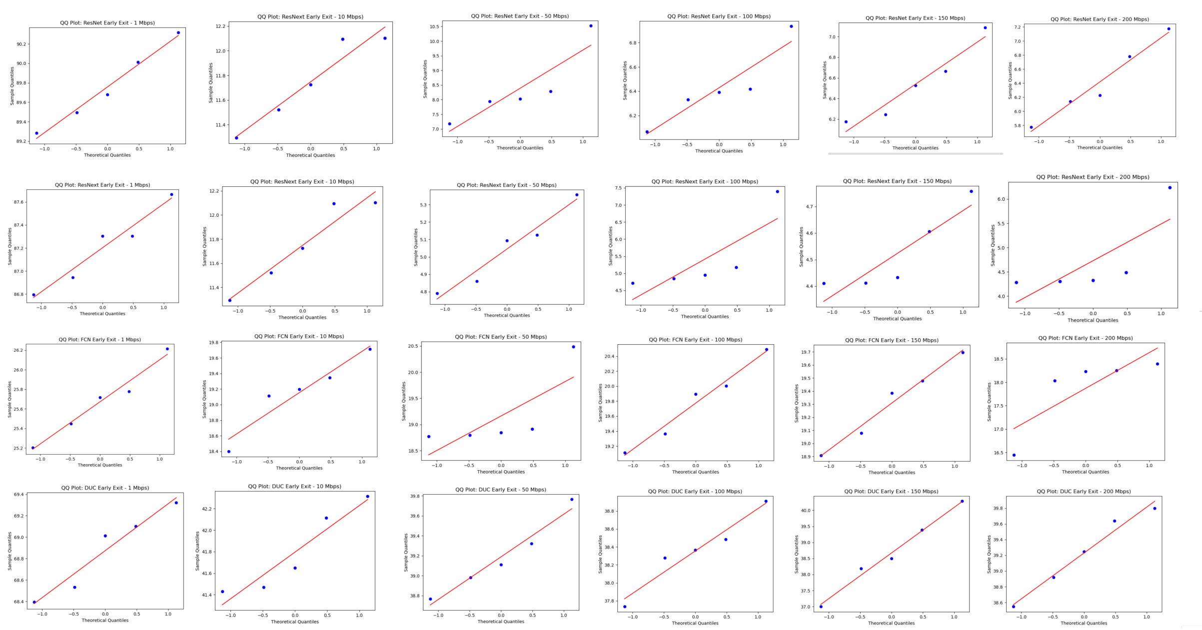


Fig. 37: Graphical Illustration of QQ plots for RQ6 Early Exit Mobile Deployment Strategies

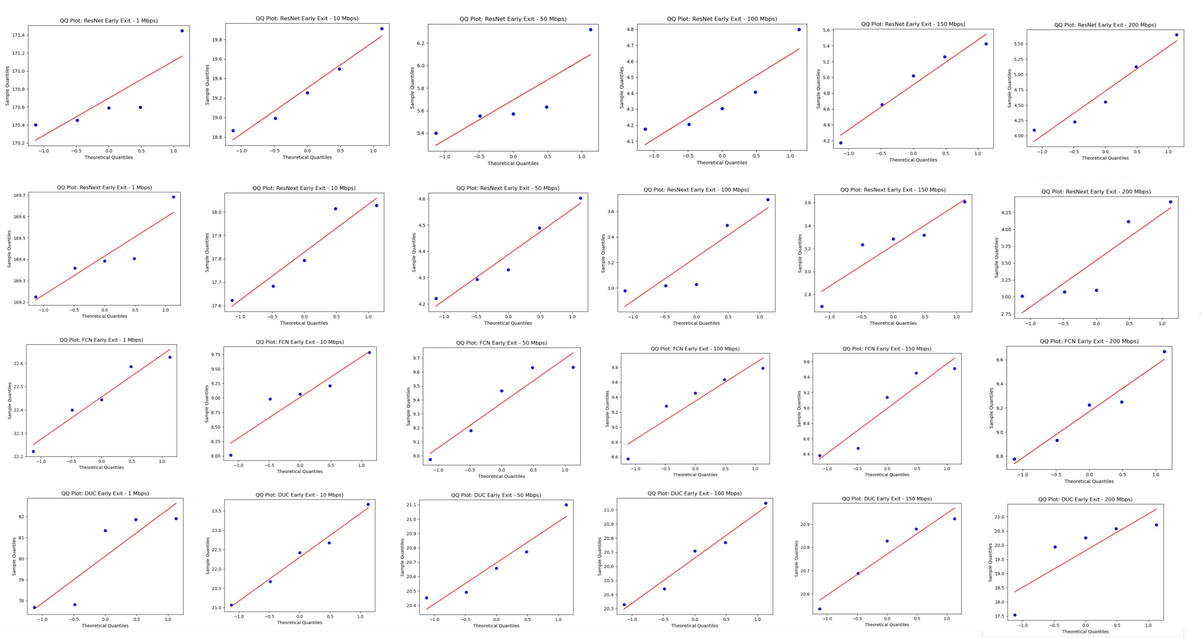


Fig. 38: Graphical Illustration of QQ plots for RQ6 Early Exit Edge Deployment Strategies

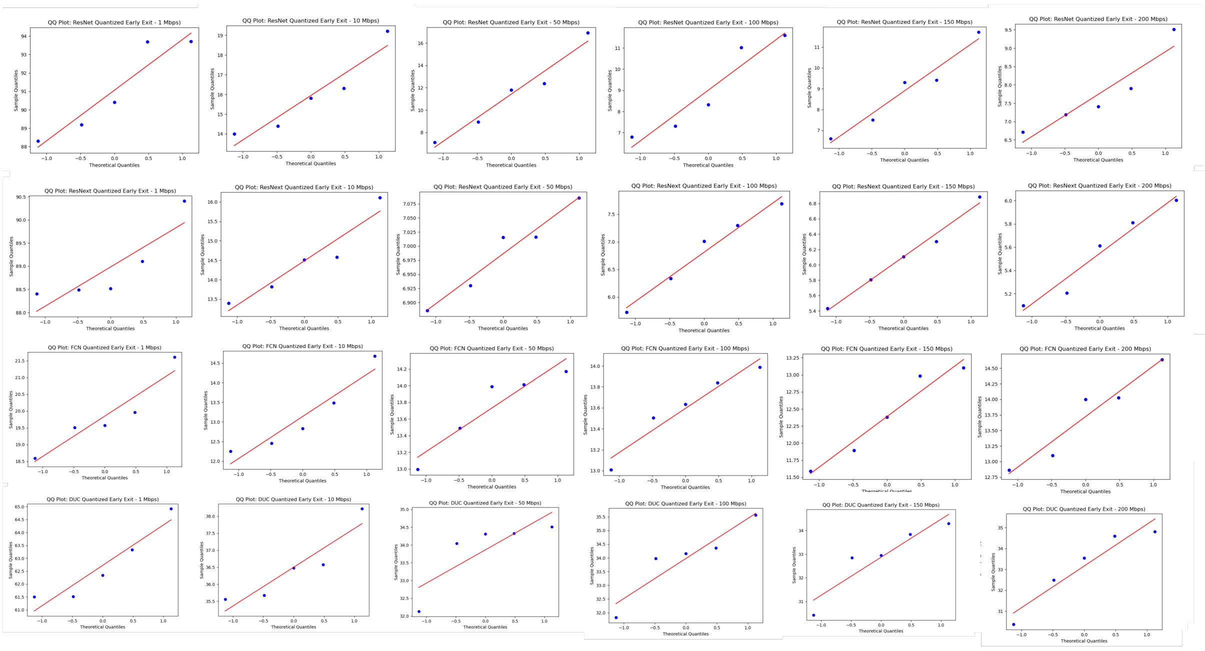


Fig. 39: Graphical Illustration of QQ plots for RQ6 Quantized Early Exit Mobile Deployment Strategies

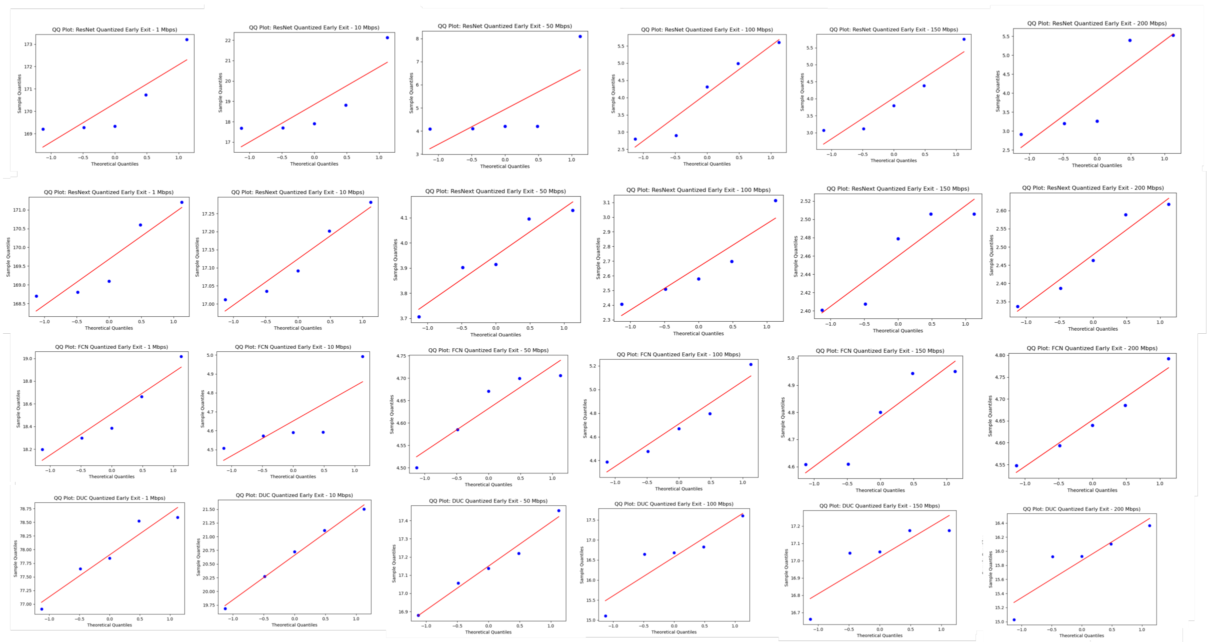


Fig. 40: Graphical Illustration of QQ plots for RQ6 Quantized Early Exit Edge Deployment Strategies

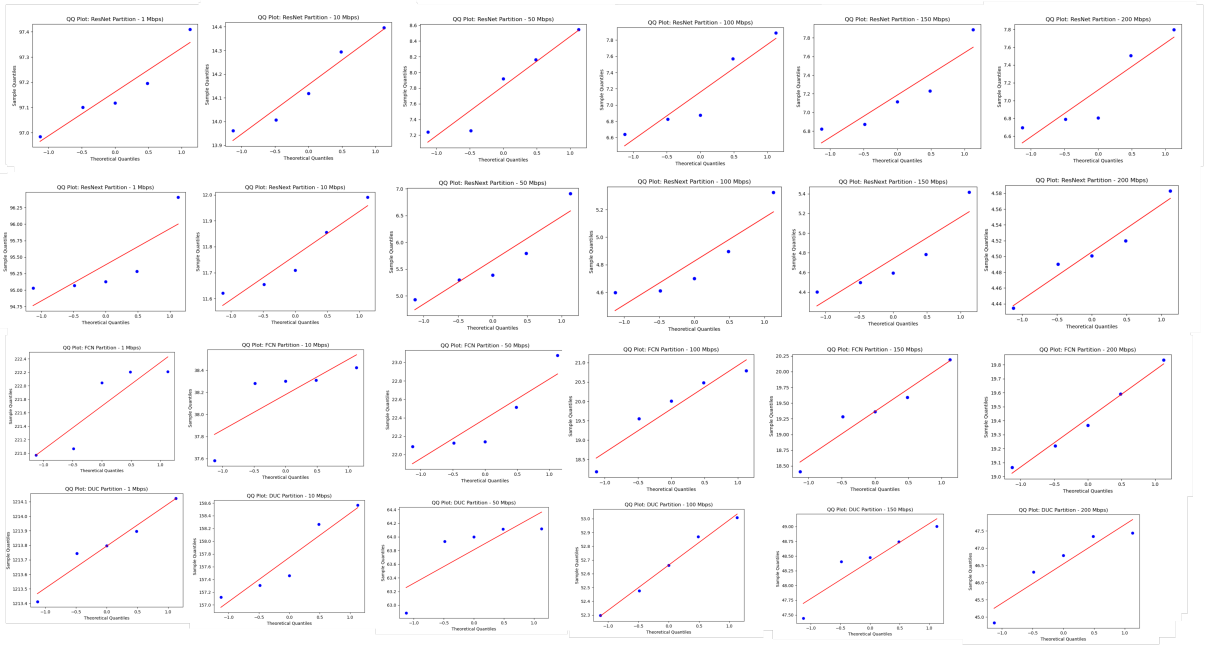


Fig. 41: Graphical Illustration of QQ plots for RQ6 Mobile-Edge Partition Deployment Strategies

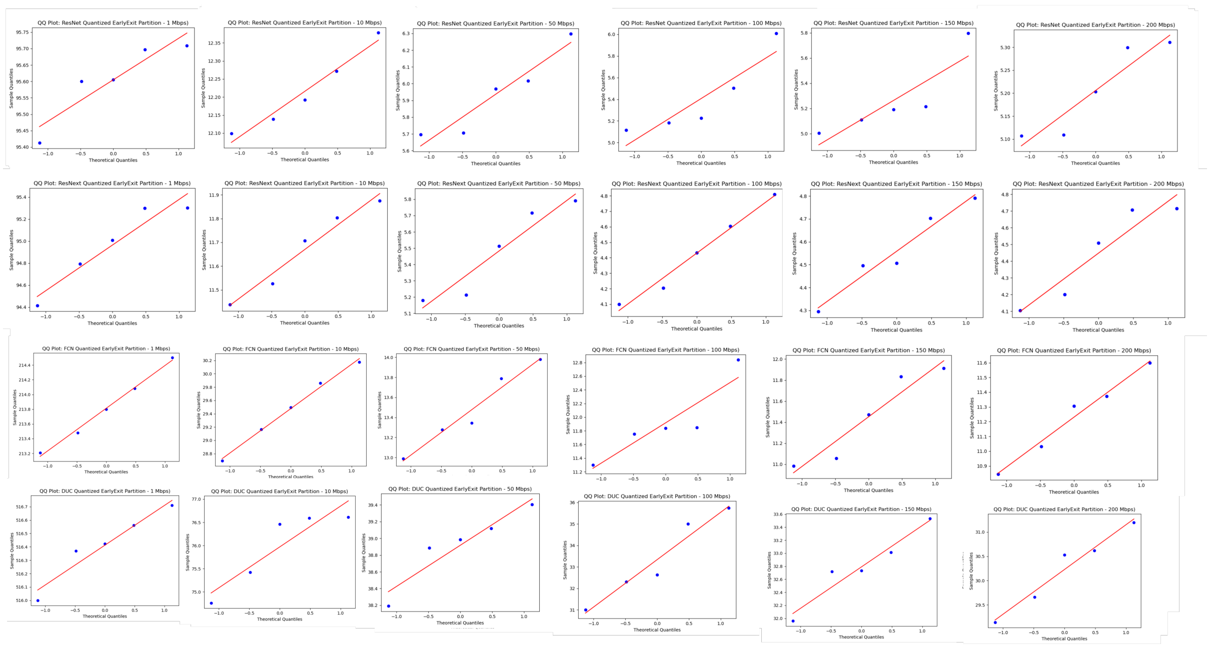


Fig. 42: Graphical Illustration of QQ plots for RQ6 Mobile-Edge Quantized Early Exit Partition Deployment Strategies

Appendix C Data Analysis and Normality Assessment

Table 26: Descriptive statistics of the latency for RQ1, including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile	ResNet	Identity	6.90	30.74	8.94	10.34	3.66	1.07×10^{-27}	No
Mobile	ResNext	Identity	4.76	41.58	6.31	8.27	5.54	3.66×10^{-33}	No
Mobile	FCN	Identity	21.91	53.66	27.39	28.78	4.73	5.30×10^{-17}	No
Mobile	DUC	Identity	42.53	47.89	45.46	45.48	0.69	5.21×10^{-5}	No
Edge	ResNet	Identity	4.89	12.90	5.64	6.10	1.28	9.59×10^{-30}	No
Edge	ResNext	Identity	3.19	45.67	3.97	5.56	4.69	8.59×10^{-36}	No
Edge	FCN	Identity	7.98	29.37	12.29	12.90	2.60	3.40×10^{-22}	No
Edge	DUC	Identity	17.93	25.11	23.65	23.05	1.50	9.40×10^{-23}	No
Cloud	ResNet	Identity	11.14	84.05	17.62	22.51	13.25	1.97×10^{-26}	No
Cloud	ResNext	Identity	11.19	83.89	17.61	22.49	13.21	2.04×10^{-26}	No
Cloud	FCN	Identity	3.24	7.83	3.77	3.89	0.55	8.64×10^{-30}	No
Cloud	DUC	Identity	27.55	31.34	28.62	28.85	0.89	5.86×10^{-15}	No

Table 27: Descriptive statistics of the latency for RQ2, including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile	ResNet	Quantize	6.76	23.09	10.67	11.37	3.34	1.88×10^{-19}	No
Mobile	ResNext	Quantize	5.22	15.49	7.29	7.92	1.89	5.74×10^{-11}	No
Mobile	FCN	Quantize	13.71	32.78	19.86	20.29	2.89	2.02×10^{-11}	No
Mobile	DUC	Quantize	36.16	41.47	38.70	38.81	1.00	2.32×10^{-7}	No
Edge	ResNet	Quantize	2.89	10.96	3.48	4.56	1.96	2.63×10^{-13}	No
Edge	ResNext	Quantize	2.19	7.02	2.67	2.83	0.66	1.50×10^{-12}	No
Edge	FCN	Quantize	5.47	16.53	7.49	7.95	1.78	6.52×10^{-9}	No
Edge	DUC	Quantize	17.15	21.42	19.59	19.50	0.65	1.79×10^{-13}	No
Cloud	ResNet	Quantize	11.16	83.96	17.64	22.49	13.21	1.98×10^{-26}	No
Cloud	ResNext	Quantize	11.17	83.85	17.66	22.49	13.21	2.04×10^{-26}	No
Cloud	FCN	Quantize	3.15	7.73	3.78	3.90	0.56	7.97×10^{-30}	No
Cloud	DUC	Quantize	29.60	36.28	31.00	31.24	1.00	2.01×10^{-14}	No

Table 28: Descriptive statistics of the latency for RQ3, including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile	ResNet	Early Exit	5.78	19.21	7.47	7.97	1.77	1.37×10^{-12}	No
Mobile	ResNext	Early Exit	4.00	22.47	5.02	5.61	2.00	1.42×10^{-12}	No
Mobile	FCN	Early Exit	17.19	48.63	26.02	26.74	4.86	2.22×10^{-9}	No
Mobile	DUC	Early Exit	38.88	46.83	40.90	40.99	0.72	7.85×10^{-8}	No
Edge	ResNet	Early Exit	3.75	5.79	4.40	4.44	0.31	6.44×10^{-7}	No
Edge	ResNext	Early Exit	2.46	4.63	2.99	3.05	0.31	1.58×10^{-6}	No
Edge	FCN	Early Exit	6.94	26.04	11.35	12.12	2.98	1.03×10^{-11}	No
Edge	DUC	EarlyExit	15.70	25.37	20.34	19.90	1.60	1.88×10^{-12}	No
Cloud	ResNet	Early Exit	11.17	83.95	17.55	22.51	13.23	1.95×10^{-26}	No
Cloud	ResNext	Early Exit	11.13	84.05	17.66	22.50	13.22	2.16×10^{-26}	No
Cloud	FCN	Early Exit	3.25	7.90	3.74	3.89	0.56	8.61×10^{-30}	No
Cloud	DUC	Early Exit	27.42	49.49	28.62	29.02	1.44	2.78×10^{-9}	No

Table 29: Descriptive statistics of the latency for RQ4, including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Edge	ResNet	Partition	6.03	27.02	8.06	8.73	2.46	3.65×10^{-11}	No
Mobile-Edge	ResNext	Partition	4.09	17.42	5.00	5.42	1.33	4.03×10^{-9}	No
Mobile-Edge	FCN	Partition	18.35	36.01	22.66	23.77	3.44	2.11×10^{-10}	No
Mobile-Edge	DUC	Partition	45.49	49.93	47.81	47.75	0.63	2.79×10^{-5}	No
Edge-Cloud	ResNet	Partition	10.58	23.93	11.10	11.33	1.10	1.02×10^{-7}	No
Edge-Cloud	ResNext	Partition	9.73	11.68	10.18	10.22	0.25	3.19×10^{-6}	No
Edge-Cloud	FCN	Partition	175.07	308.50	224.46	235.73	34.07	7.53×10^{-18}	No
Edge-Cloud	DUC	Partition	1152.53	1159.75	1155.72	1155.91	1.45	2.39×10^{-6}	No
Mobile-Cloud	ResNet	Partition	12.29	28.26	14.02	14.62	2.17	5.18×10^{-10}	No
Mobile-Cloud	ResNext	Partition	10.69	31.60	11.88	12.34	1.67	3.23×10^{-9}	No
Mobile-Cloud	FCN	Partition	186.09	323.57	235.80	247.88	35.81	1.68×10^{-11}	No
Mobile-Cloud	DUC	Partition	1172.62	1244.12	1192.89	1200.53	17.37	5.87×10^{-6}	No

Table 30: Descriptive statistics of the latency for RQ5, including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile	ResNet	Quantize Early Exit	5.69	20.66	8.29	9.09	2.67	2.12×10^{-9}	No
Edge	ResNet	Quantize Early Exit	2.39	8.34	2.89	3.20	0.94	2.98×10^{-4}	No
Cloud	ResNet	Quantize Early Exit	11.18	83.95	17.60	22.50	13.22	3.24×10^{-12}	No
Mobile	ResNext	Quantize Early Exit	4.40	14.28	6.68	7.06	1.55	1.02×10^{-6}	No
Edge	ResNext	Quantize Early Exit	1.99	5.13	2.30	2.37	0.32	2.61×10^{-4}	No
Cloud	ResNext	Quantize Early Exit	11.17	83.98	17.67	22.50	13.21	2.97×10^{-13}	No
Mobile	FCN	Quantize Early Exit	12.22	26.65	16.85	17.33	2.67	9.52×10^{-7}	No
Edge	FCN	Quantize EarlyExit	4.57	16.45	6.22	6.63	1.51	3.13×10^{-5}	No
Cloud	FCN	Quantize Early Exit	3.30	7.82	3.76	3.89	0.56	2.14×10^{-4}	No
Mobile	DUC	Quantize EarlyExit	31.67	56.66	35.21	35.44	1.58	5.72×10^{-8}	No
Edge	DUC	Quantize Early Exit	15.35	33.97	17.30	17.24	1.26	3.18×10^{-6}	No
Cloud	DUC	Quantize EarlyExit	29.06	47.22	30.70	31.06	1.77	1.51×10^{-7}	No
Mobile-Edge	ResNet	Quantize Early Exit Partition	3.99	11.51	5.63	5.84	1.13	2.34×10^{-6}	No
Edge-Cloud	ResNet	Quantize Early Exit Partition	9.37	10.53	9.76	9.76	0.17	1.79×10^{-5}	No
Mobile-Cloud	ResNet	Quantize Early Exit	11.31	18.04	12.49	12.76	1.07	5.18×10^{-7}	No
Mobile-Edge	ResNext	Quantize Early Exit Partition	3.59	10.08	4.77	5.01	0.93	3.06×10^{-7}	No
Edge-Cloud	ResNext	Quantize Early Exit Partition	9.20	10.21	9.59	9.61	0.17	1.88×10^{-5}	No
Mobile-Cloud	ResNext	Quantize Early Exit Partition	10.79	15.80	11.93	12.09	0.72	2.31×10^{-5}	No
Mobile-Edge	FCN	Quantize Early Exit Partition	10.32	23.18	14.08	14.62	2.27	7.12×10^{-6}	No
Edge-Cloud	FCN	Quantize Early Exit Partition	173.28	298.98	220.38	231.49	33.28	1.31×10^{-11}	No
Mobile-Cloud	FCN	Quantize Early Exit Partition	178.66	308.73	228.21	239.51	34.31	2.93×10^{-12}	No
Mobile-Edge	DUC	Quantize Early Exit Partition	29.33	34.06	32.01	31.80	1.12	2.17×10^{-5}	No
Edge-Cloud	DUC	Quantize Early Exit Partition	466.66	472.78	469.10	469.26	0.95	3.11×10^{-4}	No
Mobile-Cloud	DUC	Quantize Early Exit Partition	485.44	512.84	488.60	490.17	4.92	4.81×10^{-6}	No

C.1 Hypothesis Testing

Table 31: Descriptive statistics of the latency for RQ6 (Cloud Identity Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Identity	1	1	250.811300	250.973800	250.938000	250.922300	0.059400	0.140800	Yes
Cloud	ResNext	Identity	1	1	250.588000	250.790100	250.698300	250.693400	0.068000	0.997600	Yes
Cloud	FCN	Identity	1	1	20.544300	20.689900	20.585200	20.610800	0.056400	0.411200	Yes
Cloud	DUC	Identity	1	1	91.961000	92.196400	92.038100	92.057300	0.084000	0.726200	Yes
Cloud	ResNet	Identity	1	10	175.653100	175.813400	175.794100	175.755700	0.063400	0.169900	Yes
Cloud	ResNext	Identity	1	10	175.091500	175.209100	175.101900	175.125300	0.044000	0.048200	No
Cloud	FCN	Identity	1	10	14.890900	15.124900	14.919300	14.968000	0.087300	0.165300	Yes
Cloud	DUC	Identity	1	10	65.639400	65.819800	65.797600	65.747900	0.075900	0.099100	Yes
Cloud	ResNet	Identity	1	50	168.998500	169.377200	169.140100	169.160200	0.122500	0.473600	Yes
Cloud	ResNext	Identity	1	50	168.204300	168.428100	168.295400	168.305100	0.071600	0.501400	Yes
Cloud	FCN	Identity	1	50	14.882500	14.896900	14.886800	14.888800	0.005200	0.695500	Yes
Cloud	DUC	Identity	1	50	63.523800	63.643000	63.608700	63.594900	0.039900	0.532100	Yes
Cloud	ResNet	Identity	1	100	168.511600	168.821100	168.690200	168.674200	0.098900	0.650300	Yes
Cloud	ResNext	Identity	1	100	167.912900	168.085800	167.966600	167.983600	0.062700	0.673700	Yes
Cloud	FCN	Identity	1	100	14.890600	15.223300	15.005500	15.031900	0.114700	0.847000	Yes
Cloud	DUC	Identity	1	100	63.032300	63.666900	63.527700	63.431300	0.229700	0.424000	Yes
Cloud	ResNet	Identity	1	150	168.479000	168.832100	168.591800	168.603300	0.127300	0.297600	Yes
Cloud	ResNext	Identity	1	150	167.900300	168.083400	168.005800	167.998700	0.058400	0.569500	Yes
Cloud	FCN	Identity	1	150	14.869000	15.099800	14.991200	14.985100	0.095000	0.328500	Yes
Cloud	DUC	Identity	1	150	63.292700	63.840300	63.641500	63.615900	0.202900	0.625800	Yes
Cloud	ResNet	Identity	1	200	168.599100	168.940100	168.758000	168.732600	0.125300	0.348400	Yes
Cloud	ResNext	Identity	1	200	167.799900	167.976600	167.861500	167.870300	0.059400	0.577400	Yes
Cloud	FCN	Identity	1	200	14.962100	15.113300	15.051600	15.042600	0.055500	0.816600	Yes
Cloud	DUC	Identity	1	200	63.509500	63.759100	63.584100	63.613000	0.091000	0.660800	Yes
Cloud	ResNet	Identity	10	1	99.538600	99.868200	99.708700	99.687300	0.124500	0.586900	Yes
Cloud	ResNext	Identity	10	1	98.620300	100.406200	99.836300	99.542200	0.671000	0.519700	Yes
Cloud	FCN	Identity	10	1	8.086800	8.277200	8.184300	8.187000	0.061100	0.750900	Yes
Cloud	DUC	Identity	10	1	36.810900	37.232300	37.017000	37.034800	0.151400	0.864200	Yes
Cloud	ResNet	Identity	10	10	23.034800	23.188800	23.075200	23.102400	0.058600	0.460400	Yes
Cloud	ResNext	Identity	10	10	22.942300	23.035700	23.015700	23.001500	0.032300	0.310800	Yes
Cloud	FCN	Identity	10	10	0.433000	0.864500	0.731900	0.677600	0.174400	0.322700	Yes
Cloud	DUC	Identity	10	10	7.203200	7.323000	7.281200	7.274300	0.044600	0.592500	Yes
Cloud	ResNet	Identity	10	50	16.419500	16.540800	16.431000	16.456100	0.044400	0.057600	Yes
Cloud	ResNext	Identity	10	50	16.175300	16.324900	16.296300	16.260400	0.060100	0.207800	Yes
Cloud	FCN	Identity	10	50	0.452000	0.886000	0.704200	0.660900	0.167400	0.512800	Yes
Cloud	DUC	Identity	10	50	5.271800	5.382500	5.300800	5.307900	0.039400	0.118600	Yes
Cloud	ResNet	Identity	10	100	16.003300	16.172300	16.047800	16.080000	0.065200	0.390700	Yes
Cloud	ResNext	Identity	10	100	15.749100	15.977700	15.828900	15.858200	0.079400	0.836600	Yes
Cloud	FCN	Identity	10	100	0.417000	0.776700	0.696800	0.654300	0.130200	0.265200	Yes
Cloud	DUC	Identity	10	100	5.197000	5.347900	5.306800	5.282100	0.062200	0.238900	Yes
Cloud	ResNet	Identity	10	150	15.917000	16.277800	16.116600	16.120500	0.122600	0.866100	Yes
Cloud	ResNext	Identity	10	150	15.879600	16.114100	15.992700	15.994100	0.094000	0.436800	Yes
Cloud	FCN	Identity	10	150	0.440600	0.826900	0.732400	0.657200	0.162200	0.171400	Yes
Cloud	DUC	Identity	10	150	5.204800	5.370900	5.260700	5.277000	0.062000	0.658500	Yes
Cloud	ResNet	Identity	10	200	16.036000	16.139800	16.066000	16.075300	0.037400	0.526800	Yes
Cloud	ResNext	Identity	10	200	15.840300	16.081000	15.925700	15.953800	0.092200	0.556400	Yes
Cloud	FCN	Identity	10	200	0.426300	0.780900	0.694900	0.658300	0.120800	0.057400	Yes
Cloud	DUC	Identity	10	200	5.275900	5.344800	5.286700	5.297500	0.024800	0.099300	Yes
Cloud	ResNet	Identity	50	1	86.160600	86.360300	86.289200	86.272900	0.073600	0.726000	Yes
Cloud	ResNext	Identity	50	1	85.487600	87.046400	85.585100	85.857200	0.598400	0.002700	No
Cloud	FCN	Identity	50	1	8.166200	8.349400	8.295700	8.262800	0.070800	0.404000	Yes
Cloud	DUC	Identity	50	1	32.816400	33.001700	32.895200	32.897000	0.072300	0.487400	Yes
Cloud	ResNet	Identity	50	10	9.701300	11.022000	9.730000	9.984800	0.518900	0.000400	No
Cloud	ResNext	Identity	50	10	9.435900	9.621000	9.500100	9.512300	0.063500	0.785100	Yes
Cloud	FCN	Identity	50	10	0.425700	0.902000	0.637800	0.622000	0.163600	0.622500	Yes
Cloud	DUC	Identity	50	10	3.152700	3.316800	3.167400	3.203400	0.062000	0.099600	Yes
Cloud	ResNet	Identity	50	50	2.997100	3.169400	3.133000	3.101700	0.062900	0.456300	Yes
Cloud	ResNext	Identity	50	50	2.885400	2.983100	2.910600	2.920100	0.033600	0.237900	Yes
Cloud	FCN	Identity	50	50	0.460300	0.502800	0.480000	0.477700	0.016100	0.440300	Yes
Cloud	DUC	Identity	50	50	1.107000	1.139400	1.117800	1.120700	0.012000	0.668000	Yes
Cloud	ResNet	Identity	50	100	2.588300	2.752500	2.639100	2.656700	0.066200	0.322700	Yes
Cloud	ResNext	Identity	50	100	2.397100	2.472700	2.436000	2.434400	0.024200	0.766100	Yes
Cloud	FCN	Identity	50	100	0.404300	0.459600	0.413800	0.425900	0.021100	0.298200	Yes
Cloud	DUC	Identity	50	100	1.046500	1.141100	1.081100	1.080000	0.034100	0.241100	Yes
Cloud	ResNet	Identity	50	150	2.637100	2.834200	2.722500	2.725900	0.069600	0.917100	Yes
Cloud	ResNext	Identity	50	150	2.377600	2.638000	2.488200	2.500700	0.101100	0.568100	Yes
Cloud	FCN	Identity	50	150	0.405300	0.491800	0.425400	0.439900	0.033100	0.399500	Yes
Cloud	DUC	Identity	50	150	1.061100	2.376000	1.113100	1.348700	0.514100	0.000500	No
Cloud	ResNet	Identity	50	200	2.694700	2.757700	2.719200	2.717600	0.022700	0.283400	Yes
Cloud	ResNext	Identity	50	200	2.470400	2.547900	2.500800	2.504200	0.031200	0.365500	Yes
Cloud	FCN	Identity	50	200	0.451700	0.507500	0.491400	0.487300	0.019700	0.399700	Yes
Cloud	DUC	Identity	50	200	1.104300	2.450200	1.144100	1.399600	0.525600	0.000400	No

Table 32: Descriptive statistics of the latency for RQ6 (Cloud Identity Models), including Shapiro-Wilk p-values and normality assessment (Part 2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Identity	100	1	84.889400	85.448300	85.044300	85.120300	0.195200	0.689300	Yes
Cloud	ResNext	Identity	100	1	84.101500	84.396900	84.220500	84.226400	0.099500	0.803300	Yes
Cloud	FCN	Identity	100	1	8.026500	8.216100	8.160500	8.143300	0.067000	0.573800	Yes
Cloud	DUC	Identity	100	1	32.770200	32.967400	32.891900	32.884200	0.065000	0.691600	Yes
Cloud	ResNet	Identity	100	10	8.324600	8.455700	8.402600	8.395100	0.043200	0.941600	Yes
Cloud	ResNext	Identity	100	10	8.099700	8.263700	8.215700	8.190300	0.062900	0.477900	Yes
Cloud	FCN	Identity	100	10	0.461600	0.843200	0.636500	0.647200	0.163500	0.204700	Yes
Cloud	DUC	Identity	100	10	3.149200	3.202900	3.179400	3.181300	0.019000	0.542100	Yes
Cloud	ResNet	Identity	100	50	1.704100	1.948600	1.802300	1.810200	0.078600	0.387800	Yes
Cloud	ResNext	Identity	100	50	1.526900	1.625700	1.596000	1.583700	0.034800	0.759300	Yes
Cloud	FCN	Identity	100	50	0.397000	0.477600	0.431600	0.439200	0.030200	0.628900	Yes
Cloud	DUC	Identity	100	50	1.033600	1.174800	1.072100	1.082700	0.052100	0.306500	Yes
Cloud	ResNet	Identity	100	100	1.369200	1.599200	1.401200	1.435200	0.085000	0.045100	No
Cloud	ResNext	Identity	100	100	1.112100	1.261600	1.193600	1.180000	0.054900	0.645100	Yes
Cloud	FCN	Identity	100	100	0.406100	0.473800	0.445300	0.439200	0.024600	0.777100	Yes
Cloud	DUC	Identity	100	100	1.048400	1.138600	1.073200	1.084900	0.032900	0.601700	Yes
Cloud	ResNet	Identity	100	150	1.364000	1.454100	1.402500	1.405100	0.029400	0.903000	Yes
Cloud	ResNext	Identity	100	150	1.201100	1.314500	1.293300	1.266900	0.045800	0.202800	Yes
Cloud	FCN	Identity	100	150	0.406800	0.524400	0.482800	0.477400	0.041000	0.686000	Yes
Cloud	DUC	Identity	100	150	1.050100	1.098000	1.066800	1.073100	0.018500	0.497300	Yes
Cloud	ResNet	Identity	100	200	1.332800	1.537900	1.387800	1.414400	0.068800	0.461000	Yes
Cloud	ResNext	Identity	100	200	1.105900	1.304900	1.206600	1.204700	0.067500	0.999700	Yes
Cloud	FCN	Identity	100	200	0.411400	0.542800	0.447400	0.462600	0.044200	0.446300	Yes
Cloud	DUC	Identity	100	200	1.064900	1.101700	1.083500	1.085700	0.013900	0.526200	Yes
Cloud	ResNet	Identity	150	1	85.010600	85.113000	85.055900	85.054700	0.038100	0.654100	Yes
Cloud	ResNext	Identity	150	1	84.182100	86.193200	84.399500	84.691300	0.755800	0.003400	No
Cloud	FCN	Identity	150	1	8.094800	8.313700	8.185400	8.188100	0.071200	0.507800	Yes
Cloud	DUC	Identity	150	1	32.750800	32.918300	32.817700	32.827600	0.053700	0.634400	Yes
Cloud	ResNet	Identity	150	10	8.383300	8.485700	8.446900	8.439400	0.040500	0.457800	Yes
Cloud	ResNext	Identity	150	10	8.146200	8.203400	8.191400	8.184700	0.021200	0.208300	Yes
Cloud	FCN	Identity	150	10	0.419600	0.907600	0.690400	0.663200	0.210800	0.179900	Yes
Cloud	DUC	Identity	150	10	3.042000	3.227500	3.179700	3.166600	0.065000	0.089000	Yes
Cloud	ResNet	Identity	150	50	1.765400	1.928600	1.787800	1.817600	0.061400	0.154000	Yes
Cloud	ResNext	Identity	150	50	1.558600	1.717500	1.596500	1.612400	0.054500	0.055600	Yes
Cloud	FCN	Identity	150	50	0.420500	0.518400	0.435200	0.447900	0.035700	0.009200	No
Cloud	DUC	Identity	150	50	1.051000	1.087000	1.072200	1.068700	0.013700	0.607300	Yes
Cloud	ResNet	Identity	150	100	1.410000	1.606700	1.499300	1.505000	0.078100	0.474000	Yes
Cloud	ResNext	Identity	150	100	1.182500	1.289100	1.205400	1.217200	0.038500	0.177300	Yes
Cloud	FCN	Identity	150	100	0.420800	0.482300	0.446100	0.443900	0.022700	0.344300	Yes
Cloud	DUC	Identity	150	100	1.056100	1.244800	1.089600	1.114000	0.067100	0.040400	No
Cloud	ResNet	Identity	150	150	1.257400	1.579300	1.335800	1.380500	0.108500	0.328200	Yes
Cloud	ResNext	Identity	150	150	1.129000	1.249700	1.210200	1.198600	0.041400	0.819900	Yes
Cloud	FCN	Identity	150	150	0.412600	0.491500	0.425300	0.439900	0.029800	0.241200	Yes
Cloud	DUC	Identity	150	150	1.068200	1.142900	1.091500	1.097500	0.026100	0.645000	Yes
Cloud	ResNet	Identity	150	200	1.301100	1.461500	1.409900	1.394100	0.053800	0.745400	Yes
Cloud	ResNext	Identity	150	200	1.104100	1.397100	1.231000	1.248900	0.108900	0.764400	Yes
Cloud	FCN	Identity	150	200	0.397900	0.513200	0.414800	0.438600	0.043100	0.245800	Yes
Cloud	DUC	Identity	150	200	1.055100	1.115200	1.088900	1.086800	0.024000	0.441500	Yes
Cloud	ResNet	Identity	200	1	84.711000	85.007800	84.900500	84.882800	0.096400	0.451900	Yes
Cloud	ResNext	Identity	200	1	84.187900	84.317700	84.282100	84.268100	0.049100	0.434700	Yes
Cloud	FCN	Identity	200	1	8.079400	8.217000	8.138800	8.143700	0.044100	0.724500	Yes
Cloud	DUC	Identity	200	1	32.767500	33.134900	32.949000	32.931700	0.134100	0.734900	Yes
Cloud	ResNet	Identity	200	10	8.409900	8.548800	8.461200	8.467000	0.051600	0.620300	Yes
Cloud	ResNext	Identity	200	10	8.099300	8.285200	8.274200	8.221900	0.075300	0.091800	Yes
Cloud	FCN	Identity	200	10	0.443100	0.911300	0.677500	0.631900	0.174800	0.340100	Yes
Cloud	DUC	Identity	200	10	3.000800	3.215200	3.196400	3.138100	0.084000	0.141800	Yes
Cloud	ResNet	Identity	200	50	1.791400	1.835400	1.813200	1.814200	0.018200	0.292300	Yes
Cloud	ResNext	Identity	200	50	1.549800	1.680500	1.628100	1.625300	0.042500	0.479600	Yes
Cloud	FCN	Identity	200	50	0.455400	0.514900	0.463200	0.474600	0.021800	0.145200	Yes
Cloud	DUC	Identity	200	50	1.086300	1.136200	1.111900	1.109500	0.017800	0.865400	Yes
Cloud	ResNet	Identity	200	100	1.427300	1.572200	1.491000	1.494300	0.046200	0.572600	Yes
Cloud	ResNext	Identity	200	100	1.223000	1.268800	1.239500	1.242200	0.015200	0.709600	Yes
Cloud	FCN	Identity	200	100	0.457500	0.524400	0.477900	0.489200	0.025300	0.496100	Yes
Cloud	DUC	Identity	200	100	1.080700	1.151900	1.093800	1.110100	0.027200	0.309500	Yes
Cloud	ResNet	Identity	200	150	1.313400	1.467000	1.428100	1.417600	0.055200	0.125700	Yes
Cloud	ResNext	Identity	200	150	1.206500	1.393700	1.268800	1.286000	0.061400	0.612700	Yes
Cloud	FCN	Identity	200	150	0.435700	0.496200	0.455500	0.459500	0.020000	0.335500	Yes
Cloud	DUC	Identity	200	150	1.080900	1.143000	1.127100	1.118800	0.022200	0.562600	Yes
Cloud	ResNet	Identity	200	200	1.398400	1.482100	1.452800	1.450100	0.028800	0.485900	Yes
Cloud	ResNext	Identity	200	200	1.212400	1.318300	1.223300	1.239500	0.040000	0.008300	No
Cloud	FCN	Identity	200	200	0.444200	0.512100	0.479900	0.476200	0.025700	0.650600	Yes
Cloud	DUC	Identity	200	200	1.123700	1.199400	1.143200	1.151700	0.028300	0.409700	Yes

Table 33: Descriptive statistics of the latency for RQ6 (Cloud Quantized Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Quantization	1	1	250.400700	251.653200	250.571700	250.943900	0.557000	0.048400	No
Cloud	ResNext	Quantization	1	1	250.425100	250.578600	250.491600	250.497000	0.051100	0.985900	Yes
Cloud	FCN	Quantization	1	1	20.501000	20.582400	20.552100	20.545700	0.029300	0.842600	Yes
Cloud	DUC	Quantization	1	1	93.663800	93.787000	93.762100	93.743800	0.044300	0.306500	Yes
Cloud	ResNet	Quantization	1	10	174.600800	174.713400	174.684400	174.672200	0.041300	0.421600	Yes
Cloud	ResNext	Quantization	1	10	174.505500	174.711000	174.567200	174.582000	0.069300	0.289900	Yes
Cloud	FCN	Quantization	1	10	13.887100	14.094800	13.992500	13.992500	0.089500	0.192200	Yes
Cloud	DUC	Quantization	1	10	65.712400	66.502200	65.849400	65.942100	0.286000	0.024500	No
Cloud	ResNet	Quantization	1	50	167.777700	168.636900	167.811200	167.995200	0.326900	0.008400	No
Cloud	ResNext	Quantization	1	50	167.698700	167.833500	167.797400	167.780700	0.045900	0.525100	Yes
Cloud	FCN	Quantization	1	50	13.892400	14.159100	14.084900	14.049200	0.089600	0.584400	Yes
Cloud	DUC	Quantization	1	50	63.773400	64.165500	63.864700	63.905500	0.135900	0.119500	Yes
Cloud	ResNet	Quantization	1	100	167.399500	167.618600	167.472400	167.498200	0.073100	0.803900	Yes
Cloud	ResNext	Quantization	1	100	167.304200	167.447800	167.325800	167.349400	0.051200	0.076400	Yes
Cloud	FCN	Quantization	1	100	13.877900	14.094000	13.896800	13.952700	0.084200	0.140100	Yes
Cloud	DUC	Quantization	1	100	63.903200	64.766600	64.576700	64.419300	0.339100	0.314100	Yes
Cloud	ResNet	Quantization	1	150	167.328200	167.572500	167.421900	167.436100	0.090800	0.718000	Yes
Cloud	ResNext	Quantization	1	150	167.298200	167.600500	167.402600	167.401400	0.109700	0.177000	Yes
Cloud	FCN	Quantization	1	150	13.861400	14.178100	14.017400	14.027800	0.107300	0.978600	Yes
Cloud	DUC	Quantization	1	150	63.756600	63.950000	63.796800	63.833300	0.078300	0.227500	Yes
Cloud	ResNet	Quantization	1	200	167.488400	167.700100	167.582500	167.572500	0.077700	0.456500	Yes
Cloud	ResNext	Quantization	1	200	167.299100	167.499700	167.407900	167.406000	0.064100	0.658500	Yes
Cloud	FCN	Quantization	1	200	13.892800	14.069400	13.959400	13.966700	0.060700	0.824000	Yes
Cloud	DUC	Quantization	1	200	63.877700	64.959700	63.960200	64.200500	0.400200	0.054700	Yes
Cloud	ResNet	Quantization	10	1	98.367400	98.491300	98.464600	98.447900	0.042400	0.139000	Yes
Cloud	ResNext	Quantization	10	1	98.239200	98.494600	98.465200	98.420600	0.092700	0.022200	No
Cloud	FCN	Quantization	10	1	7.437300	7.959100	7.585300	7.659800	0.213300	0.268800	Yes
Cloud	DUC	Quantization	10	1	37.213700	38.172800	37.288600	37.443100	0.366200	0.001700	No
Cloud	ResNet	Quantization	10	10	22.865900	23.025200	22.943300	22.937900	0.058000	0.783500	Yes
Cloud	ResNext	Quantization	10	10	22.792600	22.928200	22.824000	22.843500	0.048300	0.459700	Yes
Cloud	FCN	Quantization	10	10	0.409500	0.837500	0.744400	0.647800	0.185100	0.105400	Yes
Cloud	DUC	Quantization	10	10	8.661900	8.953200	8.783300	8.811300	0.102100	0.862000	Yes
Cloud	ResNet	Quantization	10	50	16.099300	16.222300	16.121600	16.146800	0.045800	0.354000	Yes
Cloud	ResNext	Quantization	10	50	15.913600	16.210300	16.094600	16.078200	0.096100	0.719800	Yes
Cloud	FCN	Quantization	10	50	0.367100	0.798700	0.785800	0.650300	0.180800	0.057500	Yes
Cloud	DUC	Quantization	10	50	6.953000	7.089500	7.060800	7.029100	0.056900	0.151300	Yes
Cloud	ResNet	Quantization	10	100	15.647400	15.815200	15.724400	15.720300	0.060800	0.745900	Yes
Cloud	ResNext	Quantization	10	100	15.540100	15.757300	15.643200	15.646600	0.068900	0.595700	Yes
Cloud	FCN	Quantization	10	100	0.360800	0.801900	0.730700	0.649500	0.167900	0.234100	Yes
Cloud	DUC	Quantization	10	100	6.814500	7.079200	6.978500	6.950500	0.106200	0.381900	Yes
Cloud	ResNet	Quantization	10	150	15.679300	15.841100	15.716000	15.744900	0.058800	0.508100	Yes
Cloud	ResNext	Quantization	10	150	15.625500	15.973700	15.827500	15.826900	0.125400	0.739400	Yes
Cloud	FCN	Quantization	10	150	0.363900	0.861900	0.597200	0.640200	0.175500	0.702100	Yes
Cloud	DUC	Quantization	10	150	6.747100	7.060000	6.876000	6.887900	0.119600	0.615600	Yes
Cloud	ResNet	Quantization	10	200	15.614200	16.080600	15.898600	15.894900	0.157700	0.532400	Yes
Cloud	ResNext	Quantization	10	200	15.614600	15.927000	15.739700	15.737300	0.107400	0.502200	Yes
Cloud	FCN	Quantization	10	200	0.360000	0.855200	0.636200	0.654000	0.185000	0.492400	Yes
Cloud	DUC	Quantization	10	200	7.084300	7.280900	7.188900	7.181600	0.065600	0.992600	Yes
Cloud	ResNet	Quantization	50	1	85.000100	85.200300	85.098700	85.096000	0.064200	0.767000	Yes
Cloud	ResNext	Quantization	50	1	84.896000	85.106400	84.993800	85.015700	0.076600	0.508100	Yes
Cloud	FCN	Quantization	50	1	7.375300	7.789700	7.705600	7.645100	0.150500	0.305200	Yes
Cloud	DUC	Quantization	50	1	33.058200	33.415300	33.342400	33.275800	0.139400	0.304600	Yes
Cloud	ResNet	Quantization	50	10	9.314400	9.429500	9.374100	9.373100	0.037200	0.935600	Yes
Cloud	ResNext	Quantization	50	10	9.290100	9.403000	9.325500	9.344000	0.046700	0.232800	Yes
Cloud	FCN	Quantization	50	10	0.405900	0.861400	0.663900	0.630900	0.190700	0.253000	Yes
Cloud	DUC	Quantization	50	10	4.637100	5.036500	4.774300	4.796400	0.130800	0.237500	Yes
Cloud	ResNet	Quantization	50	50	2.701900	2.799900	2.786000	2.772900	0.035900	0.007900	No
Cloud	ResNext	Quantization	50	50	2.738900	2.766900	2.758900	2.755000	0.009800	0.718600	Yes
Cloud	FCN	Quantization	50	50	0.386500	0.408900	0.398000	0.397400	0.009300	0.274900	Yes
Cloud	DUC	Quantization	50	50	2.579800	2.964200	2.857900	2.826600	0.129600	0.089700	Yes
Cloud	ResNet	Quantization	50	100	2.224600	2.279300	2.277500	2.264800	0.020800	0.016900	No
Cloud	ResNext	Quantization	50	100	2.179500	2.344500	2.236800	2.241400	0.058300	0.459100	Yes
Cloud	FCN	Quantization	50	100	0.365000	0.416600	0.397900	0.393800	0.017600	0.913400	Yes
Cloud	DUC	Quantization	50	100	2.766300	2.904700	2.809800	2.830700	0.050300	0.675300	Yes
Cloud	ResNet	Quantization	50	150	2.260000	2.310800	2.280700	2.283000	0.019100	0.709600	Yes
Cloud	ResNext	Quantization	50	150	2.200900	2.369300	2.244200	2.283300	0.071300	0.118400	Yes
Cloud	FCN	Quantization	50	150	0.356400	0.410800	0.396700	0.386500	0.021400	0.334600	Yes
Cloud	DUC	Quantization	50	150	2.710400	3.017900	2.847600	2.879300	0.116600	0.493700	Yes
Cloud	ResNet	Quantization	50	200	2.348400	2.409800	2.389400	2.383600	0.021300	0.829600	Yes
Cloud	ResNext	Quantization	50	200	2.286500	2.367900	2.304100	2.320100	0.029600	0.475900	Yes
Cloud	FCN	Quantization	50	200	0.381600	0.403000	0.388500	0.389600	0.007700	0.490900	Yes
Cloud	DUC	Quantization	50	200	2.594100	2.773300	2.713600	2.700900	0.062900	0.723800	Yes

Table 34: Descriptive statistics of the latency for RQ6 (Cloud Quantized Models), including Shapiro-Wilk p-values and normality assessment (Part 2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Quantization	100	1	83.696400	83.836400	83.746600	83.764300	0.053100	0.574500	Yes
Cloud	ResNext	Quantization	100	1	83.590100	83.726400	83.696400	83.682300	0.047500	0.052400	Yes
Cloud	FCN	Quantization	100	1	7.440400	7.782000	7.651500	7.612800	0.143000	0.233900	Yes
Cloud	DUC	Quantization	100	1	33.015700	34.681600	33.080300	33.389100	0.647000	0.000600	No
Cloud	ResNet	Quantization	100	10	7.975800	8.249700	8.104500	8.119500	0.095200	0.936600	Yes
Cloud	ResNext	Quantization	100	10	7.991900	8.205400	7.997400	8.045200	0.081500	0.006700	No
Cloud	FCN	Quantization	100	10	0.377600	0.832600	0.600400	0.605900	0.145000	0.733100	Yes
Cloud	DUC	Quantization	100	10	4.852500	5.001000	4.865900	4.914300	0.065800	0.040000	No
Cloud	ResNet	Quantization	100	50	1.420000	1.710600	1.519000	1.537900	0.101000	0.739900	Yes
Cloud	ResNext	Quantization	100	50	1.392100	1.624100	1.609500	1.529700	0.105300	0.026400	No
Cloud	FCN	Quantization	100	50	0.386900	0.829000	0.393700	0.501100	0.170200	0.014800	No
Cloud	DUC	Quantization	100	50	2.668900	2.774100	2.756200	2.731900	0.043300	0.164200	Yes
Cloud	ResNet	Quantization	100	100	0.963900	1.176400	1.008100	1.038000	0.075100	0.272700	Yes
Cloud	ResNext	Quantization	100	100	0.930400	0.978900	0.963000	0.956800	0.021100	0.148000	Yes
Cloud	FCN	Quantization	100	100	0.365000	0.403900	0.398800	0.389000	0.015500	0.176700	Yes
Cloud	DUC	Quantization	100	100	2.643500	2.759000	2.709900	2.711900	0.043200	0.527900	Yes
Cloud	ResNet	Quantization	100	150	1.006900	1.113000	1.069100	1.058800	0.043100	0.326000	Yes
Cloud	ResNext	Quantization	100	150	0.971900	1.307100	1.116500	1.132400	0.136000	0.376100	Yes
Cloud	FCN	Quantization	100	150	0.370800	0.417400	0.385400	0.392800	0.017000	0.647700	Yes
Cloud	DUC	Quantization	100	150	2.794500	3.068500	2.921300	2.922800	0.099600	0.863000	Yes
Cloud	ResNet	Quantization	100	200	0.994200	1.221100	1.070800	1.090800	0.073800	0.490900	Yes
Cloud	ResNext	Quantization	100	200	0.966300	1.210400	0.994700	1.058900	0.096900	0.163200	Yes
Cloud	FCN	Quantization	100	200	0.366900	0.436200	0.390000	0.393400	0.024100	0.530800	Yes
Cloud	DUC	Quantization	100	200	2.654700	2.842500	2.782400	2.759400	0.073300	0.456700	Yes
Cloud	ResNet	Quantization	150	1	83.772000	84.004700	83.814500	83.844500	0.082800	0.062900	Yes
Cloud	ResNext	Quantization	150	1	83.672000	83.896100	83.695800	83.732700	0.082900	0.011400	No
Cloud	FCN	Quantization	150	1	7.620500	7.801200	7.685900	7.689400	0.061700	0.409900	Yes
Cloud	DUC	Quantization	150	1	32.968500	33.263300	33.070700	33.095600	0.095700	0.551600	Yes
Cloud	ResNet	Quantization	150	10	7.917400	8.135800	8.103500	8.070400	0.078000	0.023100	No
Cloud	ResNext	Quantization	150	10	7.995500	8.101400	8.055000	8.054100	0.041700	0.474100	Yes
Cloud	FCN	Quantization	150	10	0.364700	0.838000	0.712100	0.627200	0.194500	0.233000	Yes
Cloud	DUC	Quantization	150	10	4.472400	4.901400	4.755600	4.723700	0.139600	0.371000	Yes
Cloud	ResNet	Quantization	150	50	1.413600	1.702500	1.496900	1.537900	0.101600	0.732500	Yes
Cloud	ResNext	Quantization	150	50	1.335100	1.571000	1.539300	1.498700	0.087300	0.102300	Yes
Cloud	FCN	Quantization	150	50	0.364300	0.396900	0.373800	0.377400	0.010900	0.540100	Yes
Cloud	DUC	Quantization	150	50	2.788400	3.874700	2.867100	3.062400	0.408800	0.003600	No
Cloud	ResNet	Quantization	150	100	1.021100	1.172500	1.088400	1.083700	0.055800	0.575900	Yes
Cloud	ResNext	Quantization	150	100	0.977900	1.130100	1.033200	1.039900	0.051900	0.731800	Yes
Cloud	FCN	Quantization	150	100	0.370400	0.428000	0.394000	0.396300	0.018600	0.712800	Yes
Cloud	DUC	Quantization	150	100	2.611600	2.786500	2.677000	2.699000	0.067900	0.468500	Yes
Cloud	ResNet	Quantization	150	150	0.989100	1.094500	1.025000	1.038200	0.036400	0.853800	Yes
Cloud	ResNext	Quantization	150	150	1.003300	1.054900	1.032100	1.029100	0.022300	0.179600	Yes
Cloud	FCN	Quantization	150	150	0.361200	0.424800	0.385700	0.387200	0.021400	0.674100	Yes
Cloud	DUC	Quantization	150	150	2.746400	2.962900	2.852300	2.846600	0.070000	0.748600	Yes
Cloud	ResNet	Quantization	150	200	0.997000	1.130800	1.033000	1.054600	0.053500	0.318800	Yes
Cloud	ResNext	Quantization	150	200	0.976200	1.013200	1.008600	0.998100	0.016500	0.051000	Yes
Cloud	FCN	Quantization	150	200	0.364100	0.434500	0.396700	0.394200	0.024000	0.738400	Yes
Cloud	DUC	Quantization	150	200	2.572100	2.972900	2.877100	2.814000	0.160800	0.274700	Yes
Cloud	ResNet	Quantization	200	1	83.667300	83.901500	83.791300	83.783000	0.091800	0.528600	Yes
Cloud	ResNext	Quantization	200	1	83.704300	83.804800	83.803400	83.783700	0.039700	0.000300	No
Cloud	FCN	Quantization	200	1	7.569400	7.881600	7.591200	7.647300	0.117700	0.002200	No
Cloud	DUC	Quantization	200	1	33.077500	33.176600	33.142800	33.132300	0.037400	0.626700	Yes
Cloud	ResNet	Quantization	200	10	7.973100	8.209400	8.117700	8.100700	0.076100	0.681200	Yes
Cloud	ResNext	Quantization	200	10	8.028200	8.131400	8.071100	8.076800	0.039700	0.603400	Yes
Cloud	FCN	Quantization	200	10	0.388200	0.920000	0.615900	0.639200	0.205800	0.582600	Yes
Cloud	DUC	Quantization	200	10	4.853800	6.270600	4.968300	5.206800	0.533800	0.001400	No
Cloud	ResNet	Quantization	200	50	1.464600	1.687400	1.481100	1.534100	0.085000	0.096400	Yes
Cloud	ResNext	Quantization	200	50	1.402800	1.675200	1.457700	1.512400	0.099600	0.432100	Yes
Cloud	FCN	Quantization	200	50	0.376300	0.423800	0.383600	0.393000	0.016900	0.221000	Yes
Cloud	DUC	Quantization	200	50	2.751500	2.969200	2.870800	2.865700	0.069100	0.516400	Yes
Cloud	ResNet	Quantization	200	100	1.108900	1.180500	1.118900	1.133400	0.026600	0.218100	Yes
Cloud	ResNext	Quantization	200	100	1.017400	1.290800	1.071700	1.110200	0.096600	0.204700	Yes
Cloud	FCN	Quantization	200	100	0.380200	0.445700	0.393400	0.401500	0.023200	0.126500	Yes
Cloud	DUC	Quantization	200	100	2.732200	3.537100	2.778400	2.920300	0.309400	0.001400	No
Cloud	ResNet	Quantization	200	150	1.099700	1.200000	1.125500	1.137000	0.034500	0.392200	Yes
Cloud	ResNext	Quantization	200	150	1.023500	1.199600	1.077800	1.104600	0.064400	0.659400	Yes
Cloud	FCN	Quantization	200	150	0.382800	0.414200	0.388200	0.396000	0.012800	0.172000	Yes
Cloud	DUC	Quantization	200	150	2.734300	2.887800	2.860600	2.841800	0.054800	0.022000	No
Cloud	ResNet	Quantization	200	200	1.117500	1.197100	1.151800	1.152800	0.031800	0.401800	Yes
Cloud	ResNext	Quantization	200	200	1.033300	1.140400	1.064300	1.079300	0.039300	0.641100	Yes
Cloud	FCN	Quantization	200	200	0.372300	0.436300	0.392500	0.396000	0.021500	0.192400	Yes
Cloud	DUC	Quantization	200	200	2.783500	2.976600	2.855100	2.879100	0.066600	0.826100	Yes

Table 35: Descriptive statistics of the latency for RQ6 (Cloud Early Exit Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	EarlyExit	1	1	250.631100	250.896600	250.798300	250.784900	0.086100	0.430800	Yes
Cloud	ResNext	EarlyExit	1	1	250.479000	250.769800	250.561200	250.606500	0.112900	0.429600	Yes
Cloud	FCN	EarlyExit	1	1	20.507800	20.638900	20.554400	20.567400	0.043500	0.852800	Yes
Cloud	DUC	EarlyExit	1	1	92.176000	92.219400	92.200700	92.200200	0.014000	0.563100	Yes
Cloud	ResNet	EarlyExit	1	10	175.241800	175.521500	175.409200	175.379600	0.108700	0.490700	Yes
Cloud	ResNext	EarlyExit	1	10	174.798400	175.024400	174.998700	174.944200	0.084400	0.197700	Yes
Cloud	FCN	EarlyExit	1	10	14.770600	14.874900	14.785400	14.805800	0.037700	0.187200	Yes
Cloud	DUC	EarlyExit	1	10	65.255900	65.606800	65.539100	65.481700	0.127600	0.328800	Yes
Cloud	ResNet	EarlyExit	1	50	168.593300	168.880000	168.642300	168.691300	0.100400	0.158500	Yes
Cloud	ResNext	EarlyExit	1	50	168.136000	168.308600	168.188400	168.209500	0.058100	0.723300	Yes
Cloud	FCN	EarlyExit	1	50	14.758900	14.830400	14.807500	14.806300	0.025500	0.157500	Yes
Cloud	DUC	EarlyExit	1	50	63.138300	63.556700	63.502900	63.437100	0.151500	0.012400	No
Cloud	ResNet	EarlyExit	1	100	167.982300	168.232900	168.059600	168.092500	0.099000	0.399100	Yes
Cloud	ResNext	EarlyExit	1	100	167.746000	167.793900	167.773500	167.769300	0.017600	0.752000	Yes
Cloud	FCN	EarlyExit	1	100	14.690400	14.872500	14.785700	14.769800	0.069400	0.460700	Yes
Cloud	DUC	EarlyExit	1	100	63.060900	63.679800	63.531100	63.480900	0.226100	0.156600	Yes
Cloud	ResNet	EarlyExit	1	150	168.082500	168.322900	168.234100	168.204900	0.087100	0.750900	Yes
Cloud	ResNext	EarlyExit	1	150	167.693800	167.910200	167.701700	167.758000	0.083800	0.058800	Yes
Cloud	FCN	EarlyExit	1	150	14.621600	14.887400	14.760900	14.766200	0.105000	0.461600	Yes
Cloud	DUC	EarlyExit	1	150	63.529100	63.749500	63.695100	63.661700	0.076000	0.549400	Yes
Cloud	ResNet	EarlyExit	1	200	168.050800	169.787500	168.125700	168.442200	0.673600	0.000700	No
Cloud	ResNext	EarlyExit	1	200	167.587900	167.904300	167.761400	167.776200	0.117400	0.479500	Yes
Cloud	FCN	EarlyExit	1	200	14.691200	14.896600	14.877500	14.833700	0.077600	0.096100	Yes
Cloud	DUC	EarlyExit	1	200	63.415800	63.853100	63.611100	63.603800	0.151400	0.821400	Yes
Cloud	ResNet	EarlyExit	10	1	99.046500	99.219700	99.096300	99.122300	0.061100	0.722000	Yes
Cloud	ResNext	EarlyExit	10	1	98.540500	98.846800	98.721400	98.707300	0.126800	0.292500	Yes
Cloud	FCN	EarlyExit	10	1	7.845000	9.895300	7.987700	8.330200	0.784400	0.001000	No
Cloud	DUC	EarlyExit	10	1	36.475200	38.203300	36.734500	36.956200	0.631400	0.009700	No
Cloud	ResNet	EarlyExit	10	10	23.039700	23.194700	23.104000	23.109500	0.049700	0.566600	Yes
Cloud	ResNext	EarlyExit	10	10	22.821700	22.948500	22.898900	22.889000	0.043500	0.959600	Yes
Cloud	FCN	EarlyExit	10	10	0.452300	0.877700	0.649800	0.668400	0.174500	0.333000	Yes
Cloud	DUC	EarlyExit	10	10	7.151700	7.203300	7.175500	7.175000	0.018500	0.858600	Yes
Cloud	ResNet	EarlyExit	10	50	16.260700	16.415400	16.284700	16.305800	0.056400	0.038700	No
Cloud	ResNext	EarlyExit	10	50	16.098100	16.291800	16.234100	16.219600	0.066100	0.381500	Yes
Cloud	FCN	EarlyExit	10	50	0.415300	0.775300	0.707500	0.657700	0.133300	0.198800	Yes
Cloud	DUC	EarlyExit	10	50	5.157000	5.239800	5.186500	5.195000	0.027500	0.861900	Yes
Cloud	ResNet	EarlyExit	10	100	15.916600	16.086500	15.962600	15.978900	0.064300	0.374400	Yes
Cloud	ResNext	EarlyExit	10	100	15.690000	15.986100	15.761500	15.796300	0.101300	0.237700	Yes
Cloud	FCN	EarlyExit	10	100	0.434900	0.899700	0.690200	0.665000	0.180300	0.565100	Yes
Cloud	DUC	EarlyExit	10	100	5.128100	5.853500	5.241500	5.334100	0.263700	0.014600	No
Cloud	ResNet	EarlyExit	10	150	15.797400	16.039900	15.954700	15.944100	0.084700	0.651700	Yes
Cloud	ResNext	EarlyExit	10	150	15.810800	16.025500	15.925900	15.922900	0.068000	0.426500	Yes
Cloud	FCN	EarlyExit	10	150	0.427600	0.810000	0.700900	0.646200	0.154100	0.313100	Yes
Cloud	DUC	EarlyExit	10	150	5.170400	5.293200	5.225100	5.232200	0.044500	0.865400	Yes
Cloud	ResNet	EarlyExit	10	200	15.908000	16.242400	16.214600	16.114400	0.138600	0.089200	Yes
Cloud	ResNext	EarlyExit	10	200	15.717300	16.042600	15.843400	15.861800	0.108100	0.909000	Yes
Cloud	FCN	EarlyExit	10	200	0.432600	0.816400	0.776500	0.658200	0.165200	0.063200	Yes
Cloud	DUC	EarlyExit	10	200	5.234600	5.358800	5.318500	5.307900	0.048000	0.463000	Yes
Cloud	ResNet	EarlyExit	50	1	85.645700	86.070000	85.822900	85.862100	0.143600	0.895500	Yes
Cloud	ResNext	EarlyExit	50	1	85.218400	85.681600	85.334800	85.372600	0.160500	0.077300	Yes
Cloud	FCN	EarlyExit	50	1	7.999100	8.158700	8.081700	8.080400	0.050500	0.395100	Yes
Cloud	DUC	EarlyExit	50	1	32.500800	32.752400	32.597500	32.616500	0.088400	0.923800	Yes
Cloud	ResNet	EarlyExit	50	10	9.540300	9.616700	9.557500	9.575200	0.029900	0.271000	Yes
Cloud	ResNext	EarlyExit	50	10	9.405500	9.483800	9.459200	9.449800	0.031800	0.307600	Yes
Cloud	FCN	EarlyExit	50	10	0.398900	0.880400	0.636900	0.609500	0.166500	0.762500	Yes
Cloud	DUC	EarlyExit	50	10	3.082000	3.128300	3.085300	3.099900	0.019800	0.061100	Yes
Cloud	ResNet	EarlyExit	50	50	2.910000	3.021600	2.973000	2.967200	0.049000	0.149100	Yes
Cloud	ResNext	EarlyExit	50	50	2.786300	2.831100	2.804900	2.810500	0.017100	0.451900	Yes
Cloud	FCN	EarlyExit	50	50	0.404500	0.496100	0.458100	0.449600	0.031800	0.895500	Yes
Cloud	DUC	EarlyExit	50	50	1.025400	1.069400	1.039500	1.046300	0.016000	0.665300	Yes
Cloud	ResNet	EarlyExit	50	100	2.464000	2.598000	2.505000	2.513700	0.048300	0.468100	Yes
Cloud	ResNext	EarlyExit	50	100	2.390600	2.549300	2.424000	2.448200	0.056500	0.368700	Yes
Cloud	FCN	EarlyExit	50	100	0.365000	0.444200	0.387500	0.394400	0.026400	0.100000	Yes
Cloud	DUC	EarlyExit	50	100	0.971400	1.002600	0.984600	0.986700	0.010700	0.978900	Yes
Cloud	ResNet	EarlyExit	50	150	2.525000	3.402500	2.577300	2.745000	0.331700	0.004500	No
Cloud	ResNext	EarlyExit	50	150	2.331200	2.454900	2.363900	2.380000	0.041600	0.461000	Yes
Cloud	FCN	EarlyExit	50	150	0.385400	0.504800	0.413000	0.428100	0.044700	0.342600	Yes
Cloud	DUC	EarlyExit	50	150	0.974100	1.017600	0.977100	0.992200	0.020700	0.014800	No
Cloud	ResNet	EarlyExit	50	200	2.572600	2.656500	2.597900	2.611900	0.034100	0.268200	Yes
Cloud	ResNext	EarlyExit	50	200	2.440000	2.505000	2.478300	2.474200	0.021800	0.977300	Yes
Cloud	FCN	EarlyExit	50	200	0.401000	0.506500	0.416100	0.442400	0.040500	0.236500	Yes
Cloud	DUC	EarlyExit	50	200	0.973800	1.068500	0.991300	1.001900	0.033900	0.020500	No

Table 36: Descriptive statistics of the latency for RQ6 (Cloud Early Exit Models), including Shapiro-Wilk p-values and normality assessment (Part 2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	EarlyExit	100	1	84.374000	84.566600	84.450600	84.460800	0.078500	0.344100	Yes
Cloud	ResNext	EarlyExit	100	1	83.999900	84.193700	84.113100	84.117300	0.071700	0.475600	Yes
Cloud	FCN	EarlyExit	100	1	7.946900	8.056000	7.993000	7.994000	0.035900	0.748500	Yes
Cloud	DUC	EarlyExit	100	1	32.493600	32.713100	32.518400	32.552000	0.081500	0.008600	No
Cloud	ResNet	EarlyExit	100	10	8.222800	8.377900	8.294100	8.295800	0.052100	0.993300	Yes
Cloud	ResNext	EarlyExit	100	10	8.049500	8.206800	8.164800	8.139700	0.054900	0.666100	Yes
Cloud	FCN	EarlyExit	100	10	0.382900	0.708300	0.657300	0.606700	0.119300	0.121100	Yes
Cloud	DUC	EarlyExit	100	10	2.921000	3.112100	3.106600	3.068300	0.074000	0.001500	No
Cloud	ResNet	EarlyExit	100	50	1.601800	1.822100	1.689300	1.685800	0.077600	0.420700	Yes
Cloud	ResNext	EarlyExit	100	50	1.447600	1.633300	1.559000	1.549300	0.060400	0.836000	Yes
Cloud	FCN	EarlyExit	100	50	0.480700	0.530800	0.502700	0.502300	0.018100	0.765400	Yes
Cloud	DUC	EarlyExit	100	50	0.995900	1.042100	1.024800	1.020100	0.019200	0.267600	Yes
Cloud	ResNet	EarlyExit	100	100	1.277400	1.365600	1.323800	1.316800	0.034000	0.462500	Yes
Cloud	ResNext	EarlyExit	100	100	1.111600	1.172700	1.129700	1.138000	0.020800	0.739900	Yes
Cloud	FCN	EarlyExit	100	100	0.373500	0.438200	0.391700	0.398500	0.022500	0.567700	Yes
Cloud	DUC	EarlyExit	100	100	0.957100	0.999900	0.976900	0.977800	0.013600	0.523100	Yes
Cloud	ResNet	EarlyExit	100	150	1.233500	1.308100	1.283600	1.270000	0.028800	0.352000	Yes
Cloud	ResNext	EarlyExit	100	150	1.084900	1.201100	1.112400	1.124900	0.039700	0.082800	Yes
Cloud	FCN	EarlyExit	100	150	0.383600	0.488000	0.385800	0.423500	0.048000	0.018900	No
Cloud	DUC	EarlyExit	100	150	0.951400	1.052300	1.014300	1.009000	0.033200	0.757000	Yes
Cloud	ResNet	EarlyExit	100	200	1.210700	1.393600	1.304900	1.289700	0.068600	0.504100	Yes
Cloud	ResNext	EarlyExit	100	200	1.054600	1.172400	1.109000	1.110600	0.039000	0.990900	Yes
Cloud	FCN	EarlyExit	100	200	0.381400	0.474900	0.396200	0.420700	0.041200	0.067700	Yes
Cloud	DUC	EarlyExit	100	200	1.004700	1.017600	1.009800	1.010600	0.005400	0.284100	Yes
Cloud	ResNet	EarlyExit	150	1	84.517200	84.663700	84.604900	84.592400	0.060500	0.291200	Yes
Cloud	ResNext	EarlyExit	150	1	84.018200	84.189600	84.046600	84.069000	0.062300	0.050000	Yes
Cloud	FCN	EarlyExit	150	1	7.859000	7.993400	7.910300	7.912600	0.045700	0.594900	Yes
Cloud	DUC	EarlyExit	150	1	32.439600	32.782800	32.650600	32.653800	0.120300	0.377100	Yes
Cloud	ResNet	EarlyExit	150	10	8.221000	8.336500	8.318400	8.296900	0.041000	0.191700	Yes
Cloud	ResNext	EarlyExit	150	10	8.078800	8.165500	8.144800	8.131800	0.030400	0.523800	Yes
Cloud	FCN	EarlyExit	150	10	0.401400	0.881600	0.652200	0.633700	0.195700	0.355500	Yes
Cloud	DUC	EarlyExit	150	10	2.912200	3.117700	3.070600	3.051100	0.073000	0.124100	Yes
Cloud	ResNet	EarlyExit	150	50	1.610000	1.703200	1.647400	1.656200	0.038700	0.267600	Yes
Cloud	ResNext	EarlyExit	150	50	1.413800	1.708900	1.583300	1.577400	0.112400	0.619500	Yes
Cloud	FCN	EarlyExit	150	50	0.376800	0.438000	0.396900	0.400700	0.020200	0.310600	Yes
Cloud	DUC	EarlyExit	150	50	1.005200	1.068800	1.034100	1.036500	0.020600	0.921500	Yes
Cloud	ResNet	EarlyExit	150	100	1.230200	1.406200	1.282800	1.300800	0.059100	0.577000	Yes
Cloud	ResNext	EarlyExit	150	100	1.060500	1.456700	1.126300	1.169400	0.146100	0.013400	No
Cloud	FCN	EarlyExit	150	100	0.382900	0.432600	0.414200	0.409800	0.019700	0.437800	Yes
Cloud	DUC	EarlyExit	150	100	0.947100	1.002400	0.981200	0.977000	0.019600	0.895700	Yes
Cloud	ResNet	EarlyExit	150	150	1.203600	1.317000	1.278700	1.273400	0.038800	0.506500	Yes
Cloud	ResNext	EarlyExit	150	150	1.089300	1.333500	1.113300	1.150900	0.091900	0.003500	No
Cloud	FCN	EarlyExit	150	150	0.390600	0.498700	0.424200	0.431800	0.040200	0.484000	Yes
Cloud	DUC	EarlyExit	150	150	0.961800	1.099400	1.006900	1.019400	0.048200	0.795800	Yes
Cloud	ResNet	EarlyExit	150	200	1.227300	1.315000	1.276500	1.272300	0.034800	0.466400	Yes
Cloud	ResNext	EarlyExit	150	200	1.103500	1.163800	1.137600	1.135900	0.024000	0.437600	Yes
Cloud	FCN	EarlyExit	150	200	0.375900	0.428300	0.395200	0.398600	0.018600	0.845800	Yes
Cloud	DUC	EarlyExit	150	200	0.947800	1.022200	1.000100	0.991800	0.025600	0.710500	Yes
Cloud	ResNet	EarlyExit	200	1	84.472200	84.641500	84.508800	84.543200	0.064800	0.368800	Yes
Cloud	ResNext	EarlyExit	200	1	83.944800	84.118500	84.088500	84.068100	0.062800	0.019400	No
Cloud	FCN	EarlyExit	200	1	7.908100	8.005000	7.985200	7.972500	0.033700	0.111100	Yes
Cloud	DUC	EarlyExit	200	1	32.160500	32.563800	32.491000	32.417900	0.141300	0.272800	Yes
Cloud	ResNet	EarlyExit	200	10	8.291700	8.419200	8.325300	8.336400	0.046800	0.335200	Yes
Cloud	ResNext	EarlyExit	200	10	8.114600	8.209300	8.182600	8.176000	0.034200	0.335800	Yes
Cloud	FCN	EarlyExit	200	10	0.440300	0.923300	0.655500	0.638400	0.175500	0.583200	Yes
Cloud	DUC	EarlyExit	200	10	2.951400	3.152000	3.111700	3.093600	0.072900	0.026900	No
Cloud	ResNet	EarlyExit	200	50	1.665200	1.757800	1.728900	1.721200	0.032500	0.660600	Yes
Cloud	ResNext	EarlyExit	200	50	1.500100	1.698300	1.568000	1.584600	0.068000	0.849700	Yes
Cloud	FCN	EarlyExit	200	50	0.409100	0.495100	0.423600	0.442300	0.032400	0.314200	Yes
Cloud	DUC	EarlyExit	200	50	0.970700	1.053200	1.020200	1.010500	0.030400	0.656500	Yes
Cloud	ResNet	EarlyExit	200	100	1.285100	1.389000	1.364100	1.341000	0.040000	0.319800	Yes
Cloud	ResNext	EarlyExit	200	100	1.132500	1.206900	1.182400	1.178700	0.026200	0.520800	Yes
Cloud	FCN	EarlyExit	200	100	0.392100	0.460200	0.399000	0.418300	0.028000	0.111700	Yes
Cloud	DUC	EarlyExit	200	100	0.950000	1.019000	0.970500	0.984400	0.027000	0.296700	Yes
Cloud	ResNet	EarlyExit	200	150	1.297200	1.380700	1.318600	1.326200	0.028500	0.064800	Yes
Cloud	ResNext	EarlyExit	200	150	1.120800	1.267800	1.177700	1.188600	0.049500	0.958000	Yes
Cloud	FCN	EarlyExit	200	150	0.395200	0.471300	0.421000	0.433500	0.029800	0.364000	Yes
Cloud	DUC	EarlyExit	200	150	1.055200	1.112200	1.067700	1.075600	0.019500	0.182800	Yes
Cloud	ResNet	EarlyExit	200	200	1.337500	1.434500	1.377100	1.379100	0.034600	0.812900	Yes
Cloud	ResNext	EarlyExit	200	200	1.166900	1.220300	1.194300	1.193200	0.018300	0.997000	Yes
Cloud	FCN	EarlyExit	200	200	0.394300	0.484200	0.451800	0.439100	0.034600	0.469700	Yes
Cloud	DUC	EarlyExit	200	200	1.020600	1.065300	1.044200	1.040700	0.016700	0.599900	Yes

Table 37: Descriptive statistics of the latency for RQ6 (Cloud Quantized Early Exit Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Quantized EarlyExit	1	1	250.453800	250.601900	250.505200	250.506700	0.054400	0.309100	Yes
Cloud	ResNext	Quantized EarlyExit	1	1	250.395100	250.589000	250.480500	250.470400	0.070300	0.396700	Yes
Cloud	FCN	Quantized EarlyExit	1	1	20.481200	20.594500	20.499800	20.528500	0.048500	0.101500	Yes
Cloud	DUC	Quantized EarlyExit	1	1	93.567000	93.859000	93.765000	93.743100	0.096500	0.398900	Yes
Cloud	ResNet	Quantized EarlyExit	1	10	174.457300	174.613600	174.599600	174.554900	0.065100	0.092500	Yes
Cloud	ResNext	Quantized EarlyExit	1	10	174.382300	174.611100	174.435200	174.468500	0.083600	0.453500	Yes
Cloud	FCN	Quantized EarlyExit	1	10	13.921600	14.087400	13.961900	13.982700	0.056700	0.319200	Yes
Cloud	DUC	Quantized EarlyExit	1	10	65.739000	66.422600	66.176000	66.099000	0.256600	0.650400	Yes
Cloud	ResNet	Quantized EarlyExit	1	50	167.660500	167.865300	167.789600	167.781800	0.068500	0.642900	Yes
Cloud	ResNext	Quantized EarlyExit	1	50	167.616100	167.899600	167.762700	167.748600	0.096700	0.936900	Yes
Cloud	FCN	Quantized EarlyExit	1	50	13.880900	13.992400	13.991400	13.948400	0.053300	0.009900	No
Cloud	DUC	Quantized EarlyExit	1	50	63.760900	64.512600	64.033500	64.103900	0.287100	0.595600	Yes
Cloud	ResNet	Quantized EarlyExit	1	100	167.272300	167.619900	167.410900	167.411800	0.125100	0.618000	Yes
Cloud	ResNext	Quantized EarlyExit	1	100	167.195900	167.487700	167.374200	167.373900	0.098900	0.418100	Yes
Cloud	FCN	Quantized EarlyExit	1	100	13.874200	14.086900	13.965500	13.964000	0.075000	0.773200	Yes
Cloud	DUC	Quantized EarlyExit	1	100	63.780300	64.570700	63.948600	64.105000	0.311300	0.297000	Yes
Cloud	ResNet	Quantized EarlyExit	1	150	167.273600	167.532100	167.427600	167.402400	0.102200	0.436900	Yes
Cloud	ResNext	Quantized EarlyExit	1	150	167.174200	167.403400	167.315100	167.312200	0.085900	0.508600	Yes
Cloud	FCN	Quantized EarlyExit	1	150	13.886100	14.090200	13.888200	13.943800	0.079700	0.037300	No
Cloud	DUC	Quantized EarlyExit	1	150	63.789500	64.608000	64.072900	64.090800	0.286600	0.371700	Yes
Cloud	ResNet	Quantized EarlyExit	1	200	167.337600	167.911400	167.514300	167.555200	0.191500	0.259100	Yes
Cloud	ResNext	Quantized EarlyExit	1	200	167.388800	167.543800	167.407800	167.430100	0.057900	0.014900	No
Cloud	FCN	Quantized EarlyExit	1	200	13.888500	14.080000	13.987400	13.988300	0.060800	0.503700	Yes
Cloud	DUC	Quantized EarlyExit	1	200	63.823700	65.379000	64.056400	64.300900	0.571200	0.101200	Yes
Cloud	ResNet	Quantized EarlyExit	10	1	98.106600	98.433100	98.299300	98.278700	0.106800	0.925400	Yes
Cloud	ResNext	Quantized EarlyExit	10	1	98.095900	98.387200	98.292500	98.266300	0.105100	0.722300	Yes
Cloud	FCN	Quantized EarlyExit	10	1	7.583000	7.711900	7.605600	7.624200	0.047900	0.165800	Yes
Cloud	DUC	Quantized EarlyExit	10	1	37.043900	37.188200	37.149900	37.135700	0.051100	0.420200	Yes
Cloud	ResNet	Quantized EarlyExit	10	10	22.766000	22.938500	22.826800	22.834800	0.065900	0.475000	Yes
Cloud	ResNext	Quantized EarlyExit	10	10	22.706400	22.837600	22.811800	22.791900	0.046400	0.255800	Yes
Cloud	FCN	Quantized EarlyExit	10	10	0.430800	0.815000	0.767200	0.655200	0.171400	0.048400	No
Cloud	DUC	Quantized EarlyExit	10	10	8.605600	8.847400	8.666100	8.703700	0.085700	0.599900	Yes
Cloud	ResNet	Quantized EarlyExit	10	50	15.970700	16.127700	16.048500	16.046700	0.060300	0.632300	Yes
Cloud	ResNext	Quantized EarlyExit	10	50	15.992900	16.118800	16.075600	16.063900	0.051400	0.308600	Yes
Cloud	FCN	Quantized EarlyExit	10	50	0.334200	0.873900	0.609700	0.616600	0.173900	0.845500	Yes
Cloud	DUC	Quantized EarlyExit	10	50	6.565900	6.855700	6.671000	6.702200	0.099400	0.906100	Yes
Cloud	ResNet	Quantized EarlyExit	10	100	15.629100	15.799800	15.663000	15.679200	0.062300	0.041600	No
Cloud	ResNext	Quantized EarlyExit	10	100	15.505400	15.871600	15.671800	15.676800	0.150100	0.336000	Yes
Cloud	FCN	Quantized EarlyExit	10	100	0.329100	0.813100	0.668400	0.634500	0.176500	0.448100	Yes
Cloud	DUC	Quantized EarlyExit	10	100	6.767300	6.926600	6.883300	6.867000	0.053200	0.192300	Yes
Cloud	ResNet	Quantized EarlyExit	10	150	15.623800	15.813900	15.707400	15.725500	0.072100	0.553300	Yes
Cloud	ResNext	Quantized EarlyExit	10	150	15.616000	15.943900	15.712100	15.757700	0.111500	0.728800	Yes
Cloud	FCN	Quantized EarlyExit	10	150	0.390600	0.805200	0.673700	0.645300	0.145000	0.586800	Yes
Cloud	DUC	Quantized EarlyExit	10	150	6.774900	6.969800	6.850400	6.852300	0.069800	0.608200	Yes
Cloud	ResNet	Quantized EarlyExit	10	200	15.672500	15.931900	15.778800	15.796900	0.105100	0.373700	Yes
Cloud	ResNext	Quantized EarlyExit	10	200	15.479400	16.045500	15.651700	15.700800	0.186200	0.170800	Yes
Cloud	FCN	Quantized EarlyExit	10	200	0.349300	0.823100	0.666700	0.644800	0.169400	0.498900	Yes
Cloud	DUC	Quantized EarlyExit	10	200	6.872800	7.026700	6.898000	6.925300	0.058800	0.221700	Yes
Cloud	ResNet	Quantized EarlyExit	50	1	84.867700	85.013300	84.968900	84.951300	0.051800	0.753100	Yes
Cloud	ResNext	Quantized EarlyExit	50	1	84.760600	84.953400	84.902700	84.882800	0.064700	0.144100	Yes
Cloud	FCN	Quantized EarlyExit	50	1	7.579300	7.890100	7.590500	7.664600	0.118200	0.027600	No
Cloud	DUC	Quantized EarlyExit	50	1	32.738100	33.563500	32.971000	33.038800	0.280600	0.259900	Yes
Cloud	ResNet	Quantized EarlyExit	50	10	9.275600	9.350300	9.345000	9.323400	0.031000	0.084600	Yes
Cloud	ResNext	Quantized EarlyExit	50	10	9.254500	9.362000	9.313700	9.314000	0.035800	0.900300	Yes
Cloud	FCN	Quantized EarlyExit	50	10	0.390000	0.807800	0.594800	0.630900	0.149900	0.540800	Yes
Cloud	DUC	Quantized EarlyExit	50	10	4.351700	4.668000	4.491700	4.505000	0.105200	0.992000	Yes
Cloud	ResNet	Quantized EarlyExit	50	50	2.703200	2.790900	2.781000	2.754000	0.039800	0.034700	No
Cloud	ResNext	Quantized EarlyExit	50	50	2.684300	2.735200	2.711300	2.710500	0.016200	0.684000	Yes
Cloud	FCN	Quantized EarlyExit	50	50	0.353400	0.396600	0.377200	0.378800	0.014800	0.590500	Yes
Cloud	DUC	Quantized EarlyExit	50	50	2.669100	2.749800	2.730100	2.712400	0.031100	0.328900	Yes
Cloud	ResNet	Quantized EarlyExit	50	100	2.191600	2.293900	2.226700	2.235600	0.033900	0.771700	Yes
Cloud	ResNext	Quantized EarlyExit	50	100	2.194800	2.269300	2.230300	2.235900	0.029500	0.385100	Yes
Cloud	FCN	Quantized EarlyExit	50	100	0.316200	0.388400	0.369600	0.355900	0.027900	0.389400	Yes
Cloud	DUC	Quantized EarlyExit	50	100	2.500400	2.653300	2.580200	2.590000	0.056600	0.495100	Yes
Cloud	ResNet	Quantized EarlyExit	50	150	2.194800	2.394800	2.354600	2.321900	0.070800	0.351100	Yes
Cloud	ResNext	Quantized EarlyExit	50	150	2.192800	2.273900	2.231600	2.227400	0.028800	0.687800	Yes
Cloud	FCN	Quantized EarlyExit	50	150	0.323700	0.397200	0.361700	0.362000	0.023400	0.635800	Yes
Cloud	DUC	Quantized EarlyExit	50	150	2.457900	2.573800	2.527900	2.518300	0.046400	0.410700	Yes
Cloud	ResNet	Quantized EarlyExit	50	200	2.234600	2.378400	2.334400	2.309900	0.058500	0.265300	Yes
Cloud	ResNext	Quantized EarlyExit	50	200	2.296500	2.354300	2.305100	2.322100	0.025400	0.055000	Yes
Cloud	FCN	Quantized EarlyExit	50	200	0.362000	0.394200	0.376500	0.376500	0.010700	0.914600	Yes
Cloud	DUC	Quantized EarlyExit	50	200	2.512600	2.576900	2.556700	2.547700	0.024400	0.548300	Yes

Table 38: Descriptive statistics of the latency for RQ6 (Cloud Quantized Early Exit Models), including Shapiro-Wilk p-values and normality assessment (Part 2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Cloud	ResNet	Quantized EarlyExit	100	1	83.645500	83.721600	83.691200	83.690100	0.028900	0.538600	Yes
Cloud	ResNext	Quantized EarlyExit	100	1	83.566600	83.630000	83.602100	83.602800	0.024200	0.546300	Yes
Cloud	FCN	Quantized EarlyExit	100	1	7.458900	7.951000	7.589300	7.647700	0.168300	0.481500	Yes
Cloud	DUC	Quantized EarlyExit	100	1	32.855900	33.045500	32.878600	32.906100	0.070900	0.012900	No
Cloud	ResNet	Quantized EarlyExit	100	10	8.016300	8.116500	8.018100	8.040700	0.038500	0.005400	No
Cloud	ResNext	Quantized EarlyExit	100	10	8.012700	8.087900	8.072100	8.057400	0.028700	0.372200	Yes
Cloud	FCN	Quantized EarlyExit	100	10	0.331300	0.792300	0.619100	0.624800	0.168000	0.310300	Yes
Cloud	DUC	Quantized EarlyExit	100	10	4.551200	4.773500	4.673500	4.672700	0.071200	0.566600	Yes
Cloud	ResNet	Quantized EarlyExit	100	50	1.407200	1.608700	1.594800	1.529500	0.091200	0.031300	No
Cloud	ResNext	Quantized EarlyExit	100	50	1.333600	1.687600	1.536000	1.514400	0.136300	0.600300	Yes
Cloud	FCN	Quantized EarlyExit	100	50	0.389300	0.632900	0.405300	0.455200	0.091600	0.021600	No
Cloud	DUC	Quantized EarlyExit	100	50	2.593800	2.754300	2.718200	2.698100	0.058400	0.325200	Yes
Cloud	ResNet	Quantized EarlyExit	100	100	0.974300	1.055400	1.015400	1.011200	0.028900	0.831300	Yes
Cloud	ResNext	Quantized EarlyExit	100	100	0.965600	1.016200	0.998800	0.995000	0.018000	0.781100	Yes
Cloud	FCN	Quantized EarlyExit	100	100	0.327600	0.376700	0.360200	0.353800	0.020300	0.266400	Yes
Cloud	DUC	Quantized EarlyExit	100	100	2.577600	2.731500	2.593900	2.640500	0.067400	0.054400	Yes
Cloud	ResNet	Quantized EarlyExit	100	150	0.977000	1.163200	0.999500	1.026100	0.069000	0.004100	No
Cloud	ResNext	Quantized EarlyExit	100	150	0.970500	1.102800	1.011900	1.016300	0.046600	0.209200	Yes
Cloud	FCN	Quantized EarlyExit	100	150	0.329300	0.400800	0.384200	0.376100	0.024400	0.058600	Yes
Cloud	DUC	Quantized EarlyExit	100	150	2.425400	2.540300	2.491700	2.489800	0.038400	0.900800	Yes
Cloud	ResNet	Quantized EarlyExit	100	200	0.975700	1.158700	1.002500	1.028700	0.066100	0.016300	No
Cloud	ResNext	Quantized EarlyExit	100	200	0.881100	1.081800	0.977500	0.967100	0.069000	0.706800	Yes
Cloud	FCN	Quantized EarlyExit	100	200	0.328600	0.391700	0.346500	0.355200	0.021700	0.709400	Yes
Cloud	DUC	Quantized EarlyExit	100	200	2.369000	2.605200	2.471200	2.483800	0.091500	0.581100	Yes
Cloud	ResNet	Quantized EarlyExit	150	1	83.632700	83.783100	83.702900	83.707200	0.048200	0.828000	Yes
Cloud	ResNext	Quantized EarlyExit	150	1	83.604500	83.879400	83.733100	83.760800	0.101700	0.488800	Yes
Cloud	FCN	Quantized EarlyExit	150	1	7.586200	7.785500	7.598700	7.637800	0.074900	0.008400	No
Cloud	DUC	Quantized EarlyExit	150	1	32.820200	32.957200	32.909800	32.899300	0.047600	0.832200	Yes
Cloud	ResNet	Quantized EarlyExit	150	10	7.999600	8.140900	8.076600	8.072500	0.050200	0.948300	Yes
Cloud	ResNext	Quantized EarlyExit	150	10	7.986300	8.078300	8.020500	8.026400	0.034300	0.688300	Yes
Cloud	FCN	Quantized EarlyExit	150	10	0.340900	0.872400	0.618200	0.639100	0.182200	0.824800	Yes
Cloud	DUC	Quantized EarlyExit	150	10	4.349400	4.597600	4.481000	4.485500	0.088000	0.882800	Yes
Cloud	ResNet	Quantized EarlyExit	150	50	1.396800	1.611000	1.597200	1.525300	0.097000	0.028800	No
Cloud	ResNext	Quantized EarlyExit	150	50	1.344600	1.679300	1.548400	1.532000	0.109500	0.801700	Yes
Cloud	FCN	Quantized EarlyExit	150	50	0.331500	0.383600	0.362100	0.357600	0.017900	0.939300	Yes
Cloud	DUC	Quantized EarlyExit	150	50	2.656000	2.862100	2.684000	2.729300	0.077400	0.231900	Yes
Cloud	ResNet	Quantized EarlyExit	150	100	0.988300	1.200300	1.022700	1.058500	0.076700	0.186700	Yes
Cloud	ResNext	Quantized EarlyExit	150	100	0.926600	0.991700	0.985100	0.968700	0.025400	0.158400	Yes
Cloud	FCN	Quantized EarlyExit	150	100	0.344100	0.416000	0.397200	0.386600	0.025500	0.634000	Yes
Cloud	DUC	Quantized EarlyExit	150	100	2.560200	2.605900	2.573500	2.582900	0.017700	0.301600	Yes
Cloud	ResNet	Quantized EarlyExit	150	150	0.996500	1.092700	1.032000	1.042100	0.036000	0.672200	Yes
Cloud	ResNext	Quantized EarlyExit	150	150	0.915100	1.002200	0.995800	0.973200	0.033300	0.121000	Yes
Cloud	FCN	Quantized EarlyExit	150	150	0.344300	1.202300	0.381500	0.644400	0.350800	0.076200	Yes
Cloud	DUC	Quantized EarlyExit	150	150	2.558200	2.758800	2.681100	2.672400	0.065800	0.737100	Yes
Cloud	ResNet	Quantized EarlyExit	150	200	0.934800	1.028200	1.000500	0.991200	0.032000	0.627800	Yes
Cloud	ResNext	Quantized EarlyExit	150	200	0.905000	1.002000	0.959800	0.951700	0.034000	0.878100	Yes
Cloud	FCN	Quantized EarlyExit	150	200	0.363300	0.408600	0.383900	0.383300	0.016900	0.687000	Yes
Cloud	DUC	Quantized EarlyExit	150	200	2.507500	2.877700	2.659900	2.710500	0.139700	0.425800	Yes
Cloud	ResNet	Quantized EarlyExit	200	1	83.585800	83.759600	83.689100	83.684600	0.057700	0.820800	Yes
Cloud	ResNext	Quantized EarlyExit	200	1	83.602900	83.763600	83.697700	83.690700	0.051300	0.487100	Yes
Cloud	FCN	Quantized EarlyExit	200	1	7.564800	7.783200	7.593200	7.643000	0.081000	0.229600	Yes
Cloud	DUC	Quantized EarlyExit	200	1	32.787800	33.072700	32.894100	32.889600	0.103500	0.263600	Yes
Cloud	ResNet	Quantized EarlyExit	200	10	8.082400	8.184800	8.094000	8.109000	0.038300	0.005500	No
Cloud	ResNext	Quantized EarlyExit	200	10	8.001300	8.338600	8.079400	8.112300	0.117200	0.065400	Yes
Cloud	FCN	Quantized EarlyExit	200	10	0.382100	0.791800	0.607500	0.629300	0.152000	0.420400	Yes
Cloud	DUC	Quantized EarlyExit	200	10	4.627000	4.772100	4.750300	4.722900	0.055000	0.212700	Yes
Cloud	ResNet	Quantized EarlyExit	200	50	1.409200	1.609400	1.459700	1.505700	0.083600	0.143800	Yes
Cloud	ResNext	Quantized EarlyExit	200	50	1.367600	1.618400	1.440800	1.494300	0.102100	0.186900	Yes
Cloud	FCN	Quantized EarlyExit	200	50	0.378200	0.393000	0.389000	0.386900	0.005600	0.510900	Yes
Cloud	DUC	Quantized EarlyExit	200	50	2.379600	2.563300	2.551100	2.503500	0.072700	0.100500	Yes
Cloud	ResNet	Quantized EarlyExit	200	100	1.086100	1.129700	1.096600	1.104500	0.015800	0.546200	Yes
Cloud	ResNext	Quantized EarlyExit	200	100	1.016500	1.160800	1.102500	1.084000	0.051600	0.643600	Yes
Cloud	FCN	Quantized EarlyExit	200	100	0.358600	0.396000	0.364500	0.373200	0.015000	0.191100	Yes
Cloud	DUC	Quantized EarlyExit	200	100	2.548500	2.762300	2.677500	2.644200	0.083600	0.284800	Yes
Cloud	ResNet	Quantized EarlyExit	200	150	1.095100	1.755100	1.118900	1.251400	0.253700	0.002700	No
Cloud	ResNext	Quantized EarlyExit	200	150	1.073100	1.100800	1.082300	1.084500	0.009000	0.369100	Yes
Cloud	FCN	Quantized EarlyExit	200	150	0.367300	0.429400	0.394000	0.394200	0.020900	0.851300	Yes
Cloud	DUC	Quantized EarlyExit	200	150	2.661000	2.761900	2.741700	2.716200	0.042500	0.117900	Yes
Cloud	ResNet	Quantized EarlyExit	200	200	1.052500	1.108800	1.103000	1.091600	0.020500	0.060400	Yes
Cloud	ResNext	Quantized EarlyExit	200	200	1.066200	1.209900	1.082900	1.106100	0.053100	0.023200	No
Cloud	FCN	Quantized EarlyExit	200	200	0.357500	0.415200	0.369300	0.374300	0.021200	0.041500	No
Cloud	DUC	Quantized EarlyExit	200	200	2.535900	2.796000	2.655300	2.653900	0.085100	0.860500	Yes

Table 39: Descriptive statistics of the latency for RQ6 (Mobile-Cloud Partitioning Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Cloud	ResNet	Partition	1	1	102.916000	106.020500	103.488400	103.799800	1.133300	0.018800	No
Mobile-Cloud	ResNext	Partition	1	1	101.801800	103.981500	102.142200	102.441900	0.781400	0.010700	No
Mobile-Cloud	FCN	Partition	1	1	414.219500	414.745900	414.467100	414.446800	0.177200	0.730600	Yes
Mobile-Cloud	DUC	Partition	1	1	2345.790600	2348.402400	2347.423700	2347.469400	0.950400	0.259300	Yes
Mobile-Cloud	ResNet	Partition	1	10	95.938800	97.674100	96.392500	96.576900	0.593800	0.339400	Yes
Mobile-Cloud	ResNext	Partition	1	10	94.309200	96.202800	94.839300	95.019500	0.663600	0.497800	Yes
Mobile-Cloud	FCN	Partition	1	10	236.611100	237.865300	236.801100	237.089300	0.499700	0.198900	Yes
Mobile-Cloud	DUC	Partition	1	10	1320.158700	1321.394500	1320.750300	1320.794000	0.514400	0.283100	Yes
Mobile-Cloud	ResNet	Partition	1	50	95.935100	98.195000	96.362800	96.618900	0.825100	0.078800	Yes
Mobile-Cloud	ResNext	Partition	1	50	94.391300	95.968900	94.583500	94.814600	0.588400	0.016000	No
Mobile-Cloud	FCN	Partition	1	50	221.648800	222.074500	221.893600	221.859700	0.166100	0.511800	Yes
Mobile-Cloud	DUC	Partition	1	50	1228.263200	1229.769000	1229.536100	1229.152400	0.641500	0.102500	Yes
Mobile-Cloud	ResNet	Partition	1	100	96.302300	97.760000	96.367400	96.706900	0.552800	0.031600	No
Mobile-Cloud	ResNext	Partition	1	100	94.302400	96.032700	94.545200	94.847700	0.614500	0.059800	Yes
Mobile-Cloud	FCN	Partition	1	100	219.886100	220.288000	219.991700	220.031500	0.146300	0.406100	Yes
Mobile-Cloud	DUC	Partition	1	100	1216.329400	1217.245800	1216.872700	1216.807900	0.333600	0.841500	Yes
Mobile-Cloud	ResNet	Partition	1	150	95.869200	96.437500	96.224800	96.211200	0.192800	0.589600	Yes
Mobile-Cloud	ResNext	Partition	1	150	94.517600	96.304100	94.682600	94.988500	0.663300	0.005200	No
Mobile-Cloud	FCN	Partition	1	150	218.561500	220.191800	219.396100	219.464200	0.648200	0.400000	Yes
Mobile-Cloud	DUC	Partition	1	150	1212.724800	1215.166700	1213.359800	1213.799300	0.973600	0.310700	Yes
Mobile-Cloud	ResNet	Partition	1	200	96.021600	98.944500	96.338500	97.157400	1.275700	0.063400	Yes
Mobile-Cloud	ResNext	Partition	1	200	94.589200	95.123600	94.934100	94.906000	0.206100	0.470200	Yes
Mobile-Cloud	FCN	Partition	1	200	218.780100	220.825200	219.600400	219.788400	0.862900	0.226000	Yes
Mobile-Cloud	DUC	Partition	1	200	1211.564700	1213.005200	1212.268200	1212.299600	0.548300	0.674800	Yes
Mobile-Cloud	ResNet	Partition	10	1	20.223100	20.861200	20.593400	20.595400	0.234300	0.647500	Yes
Mobile-Cloud	ResNext	Partition	10	1	18.931800	19.460000	19.079700	19.126100	0.176800	0.131400	Yes
Mobile-Cloud	FCN	Partition	10	1	228.099000	229.483600	228.767500	228.874100	0.504000	0.642900	Yes
Mobile-Cloud	DUC	Partition	10	1	1281.944500	1283.510500	1282.665900	1282.774600	0.576800	0.734400	Yes
Mobile-Cloud	ResNet	Partition	10	10	11.847900	12.668400	12.177800	12.195200	0.266800	0.489600	Yes
Mobile-Cloud	ResNext	Partition	10	10	10.825700	11.413800	11.077300	11.134000	0.203100	0.837800	Yes
Mobile-Cloud	FCN	Partition	10	10	52.209100	53.296700	52.508200	52.619200	0.365900	0.315500	Yes
Mobile-Cloud	DUC	Partition	10	10	262.632900	263.547800	262.910000	262.996900	0.331100	0.629100	Yes
Mobile-Cloud	ResNet	Partition	10	50	11.732200	13.421500	12.176500	12.419400	0.593900	0.638600	Yes
Mobile-Cloud	ResNext	Partition	10	50	10.745200	11.536600	11.270100	11.188000	0.262800	0.798900	Yes
Mobile-Cloud	FCN	Partition	10	50	35.981500	36.742600	36.104700	36.240300	0.273500	0.211300	Yes
Mobile-Cloud	DUC	Partition	10	50	171.751400	172.868100	172.386100	172.313900	0.380900	0.977600	Yes
Mobile-Cloud	ResNet	Partition	10	100	11.657600	12.950100	12.183200	12.310400	0.458000	0.852300	Yes
Mobile-Cloud	ResNext	Partition	10	100	10.618700	10.997700	10.898900	10.858600	0.127000	0.124100	Yes
Mobile-Cloud	FCN	Partition	10	100	33.629000	36.324000	35.014000	34.954300	0.867800	0.877000	Yes
Mobile-Cloud	DUC	Partition	10	100	160.411900	161.442900	160.754200	160.815600	0.359700	0.632000	Yes
Mobile-Cloud	ResNet	Partition	10	150	12.163000	12.375900	12.285300	12.274300	0.068600	0.831000	Yes
Mobile-Cloud	ResNext	Partition	10	150	10.742100	11.500300	11.003200	11.056100	0.253300	0.713500	Yes
Mobile-Cloud	FCN	Partition	10	150	33.324000	35.091600	33.922900	34.069500	0.616000	0.823500	Yes
Mobile-Cloud	DUC	Partition	10	150	156.848300	157.049600	156.926400	156.929800	0.071900	0.676100	Yes
Mobile-Cloud	ResNet	Partition	10	200	11.973300	12.734300	12.369700	12.428600	0.282400	0.390100	Yes
Mobile-Cloud	ResNext	Partition	10	200	10.665300	11.421700	10.996100	11.013400	0.240500	0.315900	Yes
Mobile-Cloud	FCN	Partition	10	200	34.128000	36.083700	34.566100	34.791300	0.677600	0.133700	Yes
Mobile-Cloud	DUC	Partition	10	200	154.858900	156.045900	155.177200	155.275600	0.410100	0.223100	Yes
Mobile-Cloud	ResNet	Partition	50	1	13.718600	16.871900	14.039500	14.574600	1.182200	0.024900	No
Mobile-Cloud	ResNext	Partition	50	1	12.234900	14.266200	12.626100	12.823100	0.739000	0.025300	No
Mobile-Cloud	FCN	Partition	50	1	213.175000	213.699500	213.657500	213.483500	0.244300	0.021400	No
Mobile-Cloud	DUC	Partition	50	1	1188.660500	1189.400900	1188.768500	1188.875400	0.266400	0.010200	No
Mobile-Cloud	ResNet	Partition	50	10	6.063400	6.411900	6.251500	6.234600	0.115500	0.973200	Yes
Mobile-Cloud	ResNext	Partition	50	10	4.678500	5.574200	4.724200	4.993100	0.368500	0.077800	Yes
Mobile-Cloud	FCN	Partition	50	10	38.130400	39.023600	38.596700	38.588300	0.330900	0.809400	Yes
Mobile-Cloud	DUC	Partition	50	10	174.142800	175.267400	174.189500	174.467400	0.430400	0.048400	No
Mobile-Cloud	ResNet	Partition	50	50	5.958000	6.595600	6.169500	6.215700	0.208200	0.272700	Yes
Mobile-Cloud	ResNext	Partition	50	50	4.494000	6.947000	4.729900	5.146700	0.907900	0.005400	No
Mobile-Cloud	FCN	Partition	50	50	22.716700	23.494600	23.197000	23.173900	0.254600	0.441300	Yes
Mobile-Cloud	DUC	Partition	50	50	87.108500	87.646000	87.385700	87.368800	0.175000	0.927600	Yes
Mobile-Cloud	ResNet	Partition	50	100	5.442600	6.339100	5.698200	5.846100	0.339800	0.503000	Yes
Mobile-Cloud	ResNext	Partition	50	100	4.046600	4.270400	4.164700	4.171400	0.081500	0.754700	Yes
Mobile-Cloud	FCN	Partition	50	100	16.577600	18.745700	18.110800	17.906500	0.734900	0.500300	Yes
Mobile-Cloud	DUC	Partition	50	100	67.286900	68.241500	67.415700	67.541000	0.354400	0.008000	No
Mobile-Cloud	ResNet	Partition	50	150	5.395300	6.113900	5.870600	5.819900	0.264300	0.648300	Yes
Mobile-Cloud	ResNext	Partition	50	150	3.877800	5.762600	4.407900	4.709300	0.657400	0.699400	Yes
Mobile-Cloud	FCN	Partition	50	150	17.790900	19.384300	18.665100	18.678000	0.555600	0.897500	Yes
Mobile-Cloud	DUC	Partition	50	150	62.766000	64.109500	63.367500	63.455400	0.449300	0.947500	Yes
Mobile-Cloud	ResNet	Partition	50	200	5.706300	6.443800	5.929100	5.971200	0.251300	0.196900	Yes
Mobile-Cloud	ResNext	Partition	50	200	4.310500	5.873000	4.747400	4.869400	0.547800	0.342500	Yes
Mobile-Cloud	FCN	Partition	50	200	20.803500	21.071900	20.886400	20.922400	0.106300	0.391700	Yes
Mobile-Cloud	DUC	Partition	50	200	61.632000	62.924000	61.924200	62.040800	0.457800	0.069200	Yes

Table 40: Descriptive statistics of the latency for RQ6 (Mobile-Cloud Partitioning Models), including Shapiro-Wilk p-values and normality assessment (Part 2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Cloud	ResNet	Partition	100	1	12.988600	14.480300	13.214300	13.525900	0.584200	0.213600	Yes
Mobile-Cloud	ResNext	Partition	100	1	11.592000	12.176900	11.859700	11.895200	0.196500	0.950400	Yes
Mobile-Cloud	FCN	Partition	100	1	208.887800	212.363100	211.248900	211.052100	1.168900	0.310900	Yes
Mobile-Cloud	DUC	Partition	100	1	1176.858800	1177.779900	1177.172800	1177.233700	0.299700	0.201000	Yes
Mobile-Cloud	ResNet	Partition	100	10	4.598200	6.102900	4.965300	5.100800	0.523700	0.112500	Yes
Mobile-Cloud	ResNext	Partition	100	10	3.356900	4.015000	3.768100	3.737400	0.213500	0.443000	Yes
Mobile-Cloud	FCN	Partition	100	10	33.835100	34.608000	34.336600	34.287700	0.269400	0.796200	Yes
Mobile-Cloud	DUC	Partition	100	10	154.320700	156.578600	155.163100	155.285000	0.760800	0.855600	Yes
Mobile-Cloud	ResNet	Partition	100	50	4.211900	5.956800	4.994900	5.072700	0.602600	0.986400	Yes
Mobile-Cloud	ResNext	Partition	100	50	3.656100	4.031300	3.742300	3.795100	0.127900	0.296300	Yes
Mobile-Cloud	FCN	Partition	100	50	18.873800	19.889000	19.516400	19.416500	0.364000	0.846900	Yes
Mobile-Cloud	DUC	Partition	100	50	66.170700	67.155800	66.668300	66.717700	0.380100	0.504500	Yes
Mobile-Cloud	ResNet	Partition	100	100	4.381400	7.574600	4.875600	5.561400	1.192400	0.285600	Yes
Mobile-Cloud	ResNext	Partition	100	100	3.359200	3.953200	3.747900	3.667700	0.232000	0.446600	Yes
Mobile-Cloud	FCN	Partition	100	100	16.670800	17.774500	17.083200	17.064000	0.398200	0.245200	Yes
Mobile-Cloud	DUC	Partition	100	100	54.580500	56.450900	56.180700	55.824700	0.666700	0.150800	Yes
Mobile-Cloud	ResNet	Partition	100	150	4.771900	6.386800	4.978300	5.359800	0.609000	0.226200	Yes
Mobile-Cloud	ResNext	Partition	100	150	3.489400	5.751400	3.847200	4.294400	0.848400	0.266000	Yes
Mobile-Cloud	FCN	Partition	100	150	16.797400	18.357800	17.559400	17.574100	0.514800	0.996300	Yes
Mobile-Cloud	DUC	Partition	100	150	51.238100	52.455600	52.200300	52.011500	0.417800	0.256400	Yes
Mobile-Cloud	ResNet	Partition	100	200	4.800800	5.325700	5.058900	5.059200	0.212000	0.405500	Yes
Mobile-Cloud	ResNext	Partition	100	200	3.519400	4.458800	4.246600	4.107300	0.325300	0.394500	Yes
Mobile-Cloud	FCN	Partition	100	200	16.369900	16.828300	16.535800	16.568100	0.186300	0.329900	Yes
Mobile-Cloud	DUC	Partition	100	200	49.297800	50.684000	50.204600	50.210200	0.506200	0.188200	Yes
Mobile-Cloud	ResNet	Partition	150	1	13.046500	14.453000	13.799600	13.754100	0.490000	0.982900	Yes
Mobile-Cloud	ResNext	Partition	150	1	11.960400	14.001100	12.205000	12.776000	0.868700	0.085800	Yes
Mobile-Cloud	FCN	Partition	150	1	210.106500	211.179400	210.762800	210.633400	0.411300	0.516100	Yes
Mobile-Cloud	DUC	Partition	150	1	1173.523600	1173.838100	1173.622800	1173.653200	0.125400	0.355600	Yes
Mobile-Cloud	ResNet	Partition	150	10	5.158200	5.482400	5.400400	5.350700	0.125000	0.409500	Yes
Mobile-Cloud	ResNext	Partition	150	10	3.578300	4.528900	3.675100	3.851400	0.351100	0.040500	No
Mobile-Cloud	FCN	Partition	150	10	33.366400	34.081600	33.459200	33.577300	0.258500	0.032000	No
Mobile-Cloud	DUC	Partition	150	10	150.510300	151.589900	150.827400	150.977200	0.425000	0.398000	Yes
Mobile-Cloud	ResNet	Partition	150	50	4.850500	5.116000	4.999900	4.999300	0.087800	0.835000	Yes
Mobile-Cloud	ResNext	Partition	150	50	3.410400	4.144800	3.858700	3.813700	0.277000	0.671400	Yes
Mobile-Cloud	FCN	Partition	150	50	17.320600	18.462800	17.974100	17.946000	0.466700	0.336200	Yes
Mobile-Cloud	DUC	Partition	150	50	62.421500	63.602800	63.206400	63.144900	0.395500	0.406200	Yes
Mobile-Cloud	ResNet	Partition	150	100	4.795500	7.023900	5.262600	5.469600	0.805800	0.056600	Yes
Mobile-Cloud	ResNext	Partition	150	100	3.365600	3.770900	3.465400	3.552200	0.156500	0.333200	Yes
Mobile-Cloud	FCN	Partition	150	100	15.301300	17.269500	16.877600	16.589700	0.698900	0.255100	Yes
Mobile-Cloud	DUC	Partition	150	100	49.664600	52.442300	51.352200	51.352200	1.017200	0.523100	Yes
Mobile-Cloud	ResNet	Partition	150	150	4.765300	5.140800	4.957000	4.976300	0.129500	0.784500	Yes
Mobile-Cloud	ResNext	Partition	150	150	3.349400	3.828300	3.740800	3.665700	0.173200	0.239300	Yes
Mobile-Cloud	FCN	Partition	150	150	14.635200	16.684000	15.857400	15.809300	0.683500	0.808200	Yes
Mobile-Cloud	DUC	Partition	150	150	46.332300	48.391300	47.505500	47.383300	0.689700	0.984500	Yes
Mobile-Cloud	ResNet	Partition	150	200	4.723600	5.470100	4.974100	5.040500	0.244900	0.623400	Yes
Mobile-Cloud	ResNext	Partition	150	200	3.467700	4.543700	3.672600	3.822800	0.374100	0.074700	Yes
Mobile-Cloud	FCN	Partition	150	200	15.725600	16.736500	16.483200	16.289500	0.384100	0.430700	Yes
Mobile-Cloud	DUC	Partition	150	200	44.461500	46.798600	46.036100	45.766900	0.806600	0.847200	Yes
Mobile-Cloud	ResNet	Partition	200	1	13.539000	14.398100	14.024800	14.040000	0.308800	0.684200	Yes
Mobile-Cloud	ResNext	Partition	200	1	11.824600	12.000200	11.927700	11.927000	0.062200	0.750100	Yes
Mobile-Cloud	FCN	Partition	200	1	209.791500	210.443600	210.078900	210.101900	0.239500	0.833900	Yes
Mobile-Cloud	DUC	Partition	200	1	1170.747500	1171.956000	1171.530200	1171.509500	0.427100	0.394100	Yes
Mobile-Cloud	ResNet	Partition	200	10	5.161900	7.594300	5.584500	6.217200	1.095600	0.053100	Yes
Mobile-Cloud	ResNext	Partition	200	10	3.904800	5.962100	4.064300	4.729400	0.907600	0.042600	No
Mobile-Cloud	FCN	Partition	200	10	35.054000	35.706200	35.445400	35.385500	0.237500	0.802900	Yes
Mobile-Cloud	DUC	Partition	200	10	147.327300	149.587100	148.977500	148.809300	0.833100	0.287600	Yes
Mobile-Cloud	ResNet	Partition	200	50	5.292400	6.252400	5.617800	5.756700	0.366200	0.516900	Yes
Mobile-Cloud	ResNext	Partition	200	50	4.111400	9.105200	4.344600	5.229200	1.940900	0.000700	No
Mobile-Cloud	FCN	Partition	200	50	19.709900	20.682200	20.458300	20.229700	0.418600	0.103400	Yes
Mobile-Cloud	DUC	Partition	200	50	60.022800	61.220900	60.861000	60.818300	0.433600	0.195300	Yes
Mobile-Cloud	ResNet	Partition	200	100	5.457900	5.683600	5.575900	5.580800	0.079300	0.914100	Yes
Mobile-Cloud	ResNext	Partition	200	100	3.716700	4.731500	4.530900	4.365500	0.355200	0.300100	Yes
Mobile-Cloud	FCN	Partition	200	100	17.941600	18.753000	18.178100	18.313500	0.310900	0.464000	Yes
Mobile-Cloud	DUC	Partition	200	100	48.972200	50.689600	49.919400	49.901500	0.592400	0.969300	Yes
Mobile-Cloud	ResNet	Partition	200	150	5.319500	6.359000	5.634900	5.700000	0.355300	0.305400	Yes
Mobile-Cloud	ResNext	Partition	200	150	3.870800	7.353000	4.788200	4.979400	1.260400	0.111000	Yes
Mobile-Cloud	FCN	Partition	200	150	17.368500	18.357200	18.074700	17.991200	0.335400	0.313800	Yes
Mobile-Cloud	DUC	Partition	200	150	45.967700	46.421200	46.382400	46.242500	0.190400	0.077100	Yes
Mobile-Cloud	ResNet	Partition	200	200	5.273800	5.724900	5.552900	5.522400	0.164800	0.799300	Yes
Mobile-Cloud	ResNext	Partition	200	200	4.186000	5.679500	4.375900	4.680500	0.569100	0.171800	Yes
Mobile-Cloud	FCN	Partition	200	200	16.608700	18.390300	18.004100	17.824600	0.624300	0.033600	No
Mobile-Cloud	DUC	Partition	200	200	43.975900	45.185700	44.495500	44.574900	0.454400	0.712500	Yes

Table 41: Descriptive statistics of the latency for RQ6 (Edge-Cloud Partitioning Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Edge-Cloud	ResNet	Partition	1	1	177.165600	177.699900	177.330900	177.401100	0.181600	0.785400	Yes
Edge-Cloud	ResNext	Partition	1	1	176.452000	176.801900	176.702600	176.641400	0.134800	0.429400	Yes
Edge-Cloud	FCN	Partition	1	1	215.169000	217.434500	216.748200	216.606800	0.802000	0.419600	Yes
Edge-Cloud	DUC	Partition	1	1	1220.037700	1220.476800	1220.454200	1220.303300	0.200400	0.024400	No
Edge-Cloud	ResNet	Partition	1	10	170.128000	170.620100	170.422800	170.400300	0.165200	0.945000	Yes
Edge-Cloud	ResNext	Partition	1	10	168.870800	169.067200	169.001200	168.974500	0.074400	0.568100	Yes
Edge-Cloud	FCN	Partition	1	10	38.821300	40.080700	39.887400	39.649200	0.457300	0.223400	Yes
Edge-Cloud	DUC	Partition	1	10	191.065900	193.559100	193.258000	192.882100	0.933100	0.024200	No
Edge-Cloud	ResNet	Partition	1	50	170.059500	171.026700	170.240100	170.421400	0.348200	0.371900	Yes
Edge-Cloud	ResNext	Partition	1	50	168.794700	169.130500	169.031900	168.987100	0.122300	0.671600	Yes
Edge-Cloud	FCN	Partition	1	50	21.598400	23.493300	23.287100	22.925000	0.714400	0.089000	Yes
Edge-Cloud	DUC	Partition	1	50	99.661700	100.495500	100.200600	100.168700	0.312100	0.506100	Yes
Edge-Cloud	ResNet	Partition	1	100	170.067400	170.288900	170.212000	170.181000	0.087700	0.385000	Yes
Edge-Cloud	ResNext	Partition	1	100	169.100800	169.221600	169.198000	169.173100	0.049500	0.194800	Yes
Edge-Cloud	FCN	Partition	1	100	21.485600	22.687000	22.084700	22.056400	0.498500	0.292100	Yes
Edge-Cloud	DUC	Partition	1	100	86.115200	90.849700	89.776300	89.258500	1.683000	0.209400	Yes
Edge-Cloud	ResNet	Partition	1	150	170.170600	170.310900	170.244100	170.239200	0.058300	0.292400	Yes
Edge-Cloud	ResNext	Partition	1	150	168.940400	169.143300	168.990400	169.012700	0.068600	0.104800	Yes
Edge-Cloud	FCN	Partition	1	150	21.088000	21.898300	21.295400	21.430300	0.304800	0.495900	Yes
Edge-Cloud	DUC	Partition	1	150	82.222300	86.129800	84.585200	84.470600	1.300100	0.781100	Yes
Edge-Cloud	ResNet	Partition	1	200	170.181900	170.477000	170.250800	170.308800	0.115700	0.328300	Yes
Edge-Cloud	ResNext	Partition	1	200	168.870800	169.186900	168.984400	169.020100	0.108600	0.891200	Yes
Edge-Cloud	FCN	Partition	1	200	19.175800	21.886700	21.408000	21.067000	0.976300	0.054000	Yes
Edge-Cloud	DUC	Partition	1	200	84.746000	85.240600	85.054700	84.981100	0.183100	0.521600	Yes
Edge-Cloud	ResNet	Partition	10	1	26.233400	26.824400	26.631000	26.607500	0.200600	0.229400	Yes
Edge-Cloud	ResNext	Partition	10	1	25.289000	25.709200	25.455600	25.464800	0.154800	0.670000	Yes
Edge-Cloud	FCN	Partition	10	1	199.549900	202.295600	201.996500	201.536100	1.014400	0.020500	No
Edge-Cloud	DUC	Partition	10	1	1151.192600	1154.354800	1152.546500	1152.891900	1.119100	0.707500	Yes
Edge-Cloud	ResNet	Partition	10	10	17.808000	18.368500	18.045200	18.080600	0.185100	0.962800	Yes
Edge-Cloud	ResNext	Partition	10	10	17.179100	17.451000	17.206700	17.258200	0.101300	0.057200	Yes
Edge-Cloud	FCN	Partition	10	10	25.241600	25.611900	25.423000	25.431800	0.130500	0.960600	Yes
Edge-Cloud	DUC	Partition	10	10	136.999600	137.774400	137.349800	137.336800	0.255300	0.712400	Yes
Edge-Cloud	ResNet	Partition	10	50	17.799900	18.362600	18.121800	18.116700	0.209100	0.700600	Yes
Edge-Cloud	ResNext	Partition	10	50	17.049600	17.240600	17.176700	17.173200	0.070000	0.291300	Yes
Edge-Cloud	FCN	Partition	10	50	8.869700	10.225500	9.632100	9.598800	0.431600	0.478000	Yes
Edge-Cloud	DUC	Partition	10	50	45.461700	46.873700	46.205400	46.189200	0.449500	0.683500	Yes
Edge-Cloud	ResNet	Partition	10	100	17.879700	19.180900	18.095200	18.335000	0.476400	0.268700	Yes
Edge-Cloud	ResNext	Partition	10	100	17.030400	17.370800	17.075500	17.154500	0.129200	0.229100	Yes
Edge-Cloud	FCN	Partition	10	100	7.584100	8.254300	7.732000	7.797200	0.235700	0.051500	Yes
Edge-Cloud	DUC	Partition	10	100	34.332800	34.780600	34.571500	34.570400	0.155300	0.978000	Yes
Edge-Cloud	ResNet	Partition	10	150	17.700400	18.596300	18.109200	18.113400	0.286300	0.591400	Yes
Edge-Cloud	ResNext	Partition	10	150	17.235300	17.590800	17.264200	17.326600	0.133800	0.008000	No
Edge-Cloud	FCN	Partition	10	150	7.536100	7.936000	7.777000	7.747000	0.129500	0.808100	Yes
Edge-Cloud	DUC	Partition	10	150	30.679900	31.248300	30.870000	30.932600	0.219500	0.518600	Yes
Edge-Cloud	ResNet	Partition	10	200	17.832400	18.272200	18.054500	18.023000	0.161300	0.602000	Yes
Edge-Cloud	ResNext	Partition	10	200	16.959500	17.139700	17.107600	17.081200	0.064000	0.122700	Yes
Edge-Cloud	FCN	Partition	10	200	7.504100	8.872000	7.589700	7.852300	0.514000	0.004200	No
Edge-Cloud	DUC	Partition	10	200	28.800800	29.080600	28.950100	28.960700	0.098900	0.744800	Yes
Edge-Cloud	ResNet	Partition	50	1	12.974300	13.830200	13.387800	13.354000	0.305500	0.802200	Yes
Edge-Cloud	ResNext	Partition	50	1	11.857200	12.839000	12.120300	12.208500	0.337700	0.269100	Yes
Edge-Cloud	FCN	Partition	50	1	201.725000	202.501100	202.248900	202.131300	0.331700	0.165800	Yes
Edge-Cloud	DUC	Partition	50	1	1145.974900	1150.362800	1148.845900	1148.519900	1.546600	0.806400	Yes
Edge-Cloud	ResNet	Partition	50	10	4.859900	5.210600	4.976700	5.000000	0.132300	0.522200	Yes
Edge-Cloud	ResNext	Partition	50	10	3.901300	4.508400	4.098100	4.117700	0.211100	0.249600	Yes
Edge-Cloud	FCN	Partition	50	10	24.835200	25.435100	25.086900	25.098800	0.195900	0.764300	Yes
Edge-Cloud	DUC	Partition	50	10	129.552300	129.699500	129.669100	129.638700	0.055200	0.396800	Yes
Edge-Cloud	ResNet	Partition	50	50	4.700700	5.083300	4.785400	4.818300	0.138000	0.069200	Yes
Edge-Cloud	ResNext	Partition	50	50	3.994000	4.676800	4.069000	4.164400	0.258600	0.004600	No
Edge-Cloud	FCN	Partition	50	50	9.117900	10.544300	9.945700	9.832700	0.571400	0.406400	Yes
Edge-Cloud	DUC	Partition	50	50	41.089700	42.598800	42.025100	42.035400	0.535100	0.383000	Yes
Edge-Cloud	ResNet	Partition	50	100	4.496400	4.832800	4.633000	4.634800	0.113400	0.668300	Yes
Edge-Cloud	ResNext	Partition	50	100	3.774600	4.200300	3.862800	3.916700	0.146700	0.044900	No
Edge-Cloud	FCN	Partition	50	100	7.463800	8.351600	7.976300	7.851700	0.339900	0.331700	Yes
Edge-Cloud	DUC	Partition	50	100	30.247100	30.853800	30.327700	30.435500	0.220100	0.089100	Yes
Edge-Cloud	ResNet	Partition	50	150	4.478300	4.889700	4.773500	4.706600	0.143700	0.694000	Yes
Edge-Cloud	ResNext	Partition	50	150	3.750300	4.424100	3.913600	3.985600	0.234000	0.234900	Yes
Edge-Cloud	FCN	Partition	50	150	7.110900	7.890100	7.435100	7.490400	0.263600	0.975400	Yes
Edge-Cloud	DUC	Partition	50	150	26.449200	27.130000	26.887700	26.853200	0.240500	0.733000	Yes
Edge-Cloud	ResNet	Partition	50	200	4.733500	5.515800	4.877800	4.988900	0.287300	0.192400	Yes
Edge-Cloud	ResNext	Partition	50	200	3.738600	3.949400	3.886200	3.876400	0.076600	0.280400	Yes
Edge-Cloud	FCN	Partition	50	200	8.030800	8.323900	8.275500	8.197300	0.127900	0.074600	Yes
Edge-Cloud	DUC	Partition	50	200	24.442100	25.076000	24.970800	24.890200	0.228800	0.026100	No

Table 42: Descriptive statistics of the latency for RQ6 (Edge-Cloud Partitioning Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Edge-Cloud	ResNet	Partition	100	1	11.814200	12.277900	12.103400	12.072300	0.156600	0.927500	Yes
Edge-Cloud	ResNext	Partition	100	1	10.581200	10.902300	10.690200	10.692800	0.114100	0.199700	Yes
Edge-Cloud	FCN	Partition	100	1	201.670200	202.270100	201.997700	201.987400	0.192600	0.797700	Yes
Edge-Cloud	DUC	Partition	100	1	1149.414100	1150.620000	1150.223600	1150.150000	0.400800	0.360600	Yes
Edge-Cloud	ResNet	Partition	100	10	3.222000	3.890300	3.431300	3.520300	0.244000	0.700800	Yes
Edge-Cloud	ResNext	Partition	100	10	2.394100	2.964700	2.589000	2.615800	0.188500	0.224500	Yes
Edge-Cloud	FCN	Partition	100	10	24.983100	25.816800	25.419900	25.357500	0.300600	0.717200	Yes
Edge-Cloud	DUC	Partition	100	10	128.297900	130.111200	129.436500	129.235500	0.640700	0.857100	Yes
Edge-Cloud	ResNet	Partition	100	50	3.277600	3.721200	3.407800	3.452800	0.147200	0.386300	Yes
Edge-Cloud	ResNext	Partition	100	50	2.458600	2.670600	2.543400	2.549600	0.072500	0.878700	Yes
Edge-Cloud	FCN	Partition	100	50	9.480300	9.887000	9.653600	9.659000	0.131400	0.590500	Yes
Edge-Cloud	DUC	Partition	100	50	40.642900	41.457800	41.140000	41.147400	0.286600	0.434000	Yes
Edge-Cloud	ResNet	Partition	100	100	3.237900	3.884400	3.471700	3.548100	0.268600	0.256000	Yes
Edge-Cloud	ResNext	Partition	100	100	2.210400	3.075800	2.453800	2.514700	0.295200	0.131200	Yes
Edge-Cloud	FCN	Partition	100	100	7.298600	8.456600	7.727700	7.835200	0.415300	0.854000	Yes
Edge-Cloud	DUC	Partition	100	100	30.225900	30.557600	30.415600	30.418100	0.125700	0.548000	Yes
Edge-Cloud	ResNet	Partition	100	150	3.225000	3.665300	3.565500	3.501800	0.152200	0.369300	Yes
Edge-Cloud	ResNext	Partition	100	150	2.307200	2.809000	2.572600	2.562700	0.159200	0.578900	Yes
Edge-Cloud	FCN	Partition	100	150	7.111500	8.116400	7.748100	7.733600	0.346900	0.445700	Yes
Edge-Cloud	DUC	Partition	100	150	25.238900	27.233500	26.655100	26.480600	0.698500	0.518900	Yes
Edge-Cloud	ResNet	Partition	100	200	3.368800	4.835000	3.731400	3.834200	0.528500	0.108400	Yes
Edge-Cloud	ResNext	Partition	100	200	2.478300	2.805600	2.624300	2.639900	0.122500	0.747800	Yes
Edge-Cloud	FCN	Partition	100	200	7.225100	7.747600	7.564900	7.475000	0.207500	0.281000	Yes
Edge-Cloud	DUC	Partition	100	200	24.494300	25.106500	24.983800	24.912100	0.214200	0.026400	No
Edge-Cloud	ResNet	Partition	150	1	11.818400	12.378900	11.924200	12.027400	0.206700	0.364000	Yes
Edge-Cloud	ResNext	Partition	150	1	10.513600	11.026800	10.706700	10.728300	0.185300	0.745700	Yes
Edge-Cloud	FCN	Partition	150	1	201.498200	202.784200	202.545400	202.243700	0.522900	0.188900	Yes
Edge-Cloud	DUC	Partition	150	1	1149.652100	1150.475400	1150.238400	1150.175200	0.278300	0.209700	Yes
Edge-Cloud	ResNet	Partition	150	10	3.276400	3.970400	3.291300	3.473300	0.269000	0.040200	No
Edge-Cloud	ResNext	Partition	150	10	2.460000	2.918500	2.481600	2.570200	0.175100	0.002400	No
Edge-Cloud	FCN	Partition	150	10	23.124400	25.967000	23.837600	24.328600	1.032600	0.573900	Yes
Edge-Cloud	DUC	Partition	150	10	129.321800	129.950300	129.660800	129.604600	0.236400	0.549800	Yes
Edge-Cloud	ResNet	Partition	150	50	3.223100	3.877500	3.308800	3.415300	0.240100	0.053400	Yes
Edge-Cloud	ResNext	Partition	150	50	2.382200	2.593900	2.520600	2.497100	0.071400	0.876400	Yes
Edge-Cloud	FCN	Partition	150	50	9.346900	9.820000	9.600100	9.576500	0.180800	0.628600	Yes
Edge-Cloud	DUC	Partition	150	50	41.021100	41.569300	41.193700	41.236400	0.184100	0.511100	Yes
Edge-Cloud	ResNet	Partition	150	100	3.298800	3.926000	3.411800	3.562300	0.242100	0.274100	Yes
Edge-Cloud	ResNext	Partition	150	100	2.525400	3.414100	2.650200	2.782500	0.328800	0.053000	Yes
Edge-Cloud	FCN	Partition	150	100	7.647200	8.277900	7.930000	7.900600	0.228200	0.506000	Yes
Edge-Cloud	DUC	Partition	150	100	30.053400	30.446800	30.209200	30.260100	0.159700	0.280200	Yes
Edge-Cloud	ResNet	Partition	150	150	3.274500	3.885600	3.372200	3.467400	0.216500	0.064400	Yes
Edge-Cloud	ResNext	Partition	150	150	2.390100	2.960100	2.512200	2.575400	0.208900	0.187900	Yes
Edge-Cloud	FCN	Partition	150	150	7.202400	7.761300	7.535500	7.498000	0.196300	0.943500	Yes
Edge-Cloud	DUC	Partition	150	150	26.246000	26.906900	26.730800	26.612700	0.292400	0.098400	Yes
Edge-Cloud	ResNet	Partition	150	200	3.224400	4.023400	3.536400	3.573100	0.258100	0.579800	Yes
Edge-Cloud	ResNext	Partition	150	200	2.374800	2.659000	2.445700	2.497900	0.110300	0.379500	Yes
Edge-Cloud	FCN	Partition	150	200	6.288200	8.279700	7.466100	7.371200	0.639100	0.654900	Yes
Edge-Cloud	DUC	Partition	150	200	23.514200	25.091800	24.861200	24.659100	0.578800	0.007600	No
Edge-Cloud	ResNet	Partition	200	1	11.701400	12.519400	12.011300	12.090100	0.289800	0.886200	Yes
Edge-Cloud	ResNext	Partition	200	1	10.543100	10.925700	10.651800	10.695700	0.138200	0.620200	Yes
Edge-Cloud	FCN	Partition	200	1	200.590900	202.387200	201.890600	201.725800	0.604600	0.212800	Yes
Edge-Cloud	DUC	Partition	200	1	1150.062600	1150.568600	1150.275000	1150.317300	0.192400	0.623900	Yes
Edge-Cloud	ResNet	Partition	200	10	3.540400	4.564600	3.659400	3.822000	0.375200	0.007300	No
Edge-Cloud	ResNext	Partition	200	10	2.472000	2.975900	2.742900	2.724300	0.168400	0.995600	Yes
Edge-Cloud	FCN	Partition	200	10	23.569200	25.553300	24.795400	24.722600	0.649600	0.564600	Yes
Edge-Cloud	DUC	Partition	200	10	128.851400	129.090400	128.924900	128.958500	0.088600	0.632600	Yes
Edge-Cloud	ResNet	Partition	200	50	3.581000	3.957300	3.636700	3.704400	0.138700	0.179100	Yes
Edge-Cloud	ResNext	Partition	200	50	2.728100	2.967400	2.765400	2.813900	0.088700	0.274900	Yes
Edge-Cloud	FCN	Partition	200	50	7.883200	10.522900	10.376100	9.826600	1.000200	0.017500	No
Edge-Cloud	DUC	Partition	200	50	41.171500	41.953000	41.259500	41.409600	0.288600	0.086300	Yes
Edge-Cloud	ResNet	Partition	200	100	3.610300	4.175000	3.674000	3.828700	0.244100	0.072800	Yes
Edge-Cloud	ResNext	Partition	200	100	2.686100	3.051900	2.710100	2.786200	0.138500	0.030400	No
Edge-Cloud	FCN	Partition	200	100	8.179700	8.689200	8.584000	8.476500	0.211900	0.171000	Yes
Edge-Cloud	DUC	Partition	200	100	30.226500	30.736000	30.541000	30.521600	0.175100	0.835700	Yes
Edge-Cloud	ResNet	Partition	200	150	3.408500	4.094900	3.595000	3.664900	0.229500	0.195900	Yes
Edge-Cloud	ResNext	Partition	200	150	2.563200	3.471400	2.872600	2.895100	0.311200	0.253500	Yes
Edge-Cloud	FCN	Partition	200	150	7.749500	8.175700	8.010800	7.992300	0.140100	0.756900	Yes
Edge-Cloud	DUC	Partition	200	150	26.481200	26.855800	26.745900	26.719600	0.131000	0.387900	Yes
Edge-Cloud	ResNet	Partition	200	200	3.501100	5.292100	3.665100	4.006300	0.659100	0.023500	No
Edge-Cloud	ResNext	Partition	200	200	2.643300	2.974300	2.724400	2.770200	0.111900	0.363400	Yes
Edge-Cloud	FCN	Partition	200	200	7.679700	8.732500	7.893300	8.036300	0.363800	0.106500	Yes
Edge-Cloud	DUC	Partition	200	200	23.036400	25.150400	24.187700	24.146300	0.736800	0.979300	Yes

Table 43: Descriptive statistics of the latency for RQ6 (Mobile-Cloud Quantized Early Exit Partitioning Models), including Shapiro-Wilk p-values and normality assessment (Part 1)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	1	102.331500	103.296800	103.018100	102.977900	0.349100	0.166100	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	1	101.765400	102.971600	102.479900	102.443200	0.392500	0.722000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	1	407.431000	408.659200	407.804800	407.905500	0.414200	0.464600	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	1	967.185700	968.940200	968.343300	968.077200	0.703600	0.322500	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	10	94.872800	95.313700	95.237400	95.137200	0.177500	0.203300	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	10	93.975200	94.600700	94.394500	94.358700	0.213800	0.541500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	10	230.248000	230.682800	230.307400	230.399600	0.162200	0.224400	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	10	553.993800	557.456500	556.336300	555.981300	1.312900	0.569800	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	50	94.703000	97.854700	94.852800	95.558700	1.190000	0.022500	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	50	94.230300	95.579600	94.839600	94.843500	0.440800	0.870800	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	50	214.066500	215.090300	214.488700	214.544900	0.431700	0.240400	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	50	518.981600	520.531000	519.456600	519.690400	0.549000	0.735600	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	100	94.702700	95.199700	94.921200	94.958300	0.200900	0.363300	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	100	94.503300	94.923100	94.756100	94.721600	0.137200	0.868000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	100	212.197400	213.305500	212.688900	212.701500	0.357100	0.728100	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	100	514.364800	516.351600	515.702200	515.464100	0.708400	0.799100	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	150	94.623800	97.278600	95.144000	95.485700	0.922400	0.033300	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	150	94.576500	95.718600	94.747900	94.920600	0.417500	0.075700	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	150	212.178300	212.704000	212.292200	212.428000	0.221100	0.110500	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	150	514.066900	515.265000	514.592300	514.558600	0.418000	0.644000	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	1	200	94.596800	98.456300	94.813000	95.511000	1.475900	0.001000	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	1	200	94.331000	95.008500	94.899100	94.816300	0.246100	0.011400	No
Mobile-Cloud	FCN	Quantized Early Exit Partition	1	200	211.597200	213.187400	212.387200	212.347300	0.523000	0.944600	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	1	200	512.180400	514.596100	513.477300	513.514000	0.951600	0.460400	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	1	18.968400	22.350500	19.273400	19.831500	1.265700	0.002400	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	1	18.994500	19.293200	19.142000	19.136200	0.113900	0.654100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	1	221.557300	222.042300	221.792000	221.784300	0.170900	0.942200	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	1	523.663200	525.189900	524.051900	524.240300	0.568500	0.447600	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	10	11.281400	12.088200	11.360500	11.580300	0.316100	0.154000	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	10	11.141400	11.933400	11.183200	11.364800	0.300300	0.039500	No
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	10	45.214600	46.254200	45.795800	45.755100	0.350000	0.995700	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	10	115.255700	117.011900	116.662200	116.427400	0.613800	0.123100	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	50	11.737600	11.953600	11.822100	11.838700	0.086200	0.427900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	50	10.761300	11.375000	11.278500	11.204000	0.225200	0.015900	No
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	50	29.970400	30.542000	30.325800	30.295400	0.192300	0.877900	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	50	77.775100	80.482300	79.054000	79.065600	1.078000	0.459000	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	100	11.448000	11.732600	11.503100	11.551100	0.099400	0.291000	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	100	10.755500	11.786500	11.269900	11.254300	0.327700	0.572500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	100	27.919600	28.929900	28.183600	28.279500	0.369800	0.368500	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	100	74.541500	76.477400	75.343900	75.597900	0.726100	0.466900	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	150	11.220500	11.967200	11.460600	11.541100	0.266200	0.781700	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	150	11.251000	11.635200	11.374100	11.411500	0.128500	0.714400	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	150	27.791800	28.541100	28.311000	28.222500	0.249100	0.665300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	150	72.263100	75.112600	73.406200	73.562600	1.022700	0.881600	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	10	200	11.226600	11.779700	11.528200	11.527500	0.182500	0.930200	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	10	200	10.668700	11.540900	11.243400	11.172300	0.283400	0.454800	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	10	200	27.420500	28.097500	27.615100	27.730000	0.269800	0.329600	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	10	200	72.115400	75.761600	73.931500	73.877800	1.255000	0.991300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	1	12.549000	13.344000	12.797700	12.847700	0.270700	0.398800	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	1	12.169600	12.485500	12.404500	12.347200	0.134000	0.161500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	1	205.492300	206.367500	206.064700	205.979000	0.343100	0.467100	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	1	484.717300	486.392500	485.471000	485.493500	0.534500	0.580200	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	10	5.154100	8.623800	5.284900	5.961500	1.334900	0.001200	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	10	4.642300	5.290200	4.996600	4.953700	0.226300	0.921300	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	10	30.027200	30.459100	30.091400	30.179000	0.157000	0.243600	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	10	82.112000	83.764400	83.351800	83.033500	0.628700	0.424400	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	50	5.095800	5.321500	5.179000	5.207900	0.081400	0.737900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	50	4.682100	5.079300	4.974100	4.918500	0.137500	0.628800	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	50	14.784000	15.295500	15.001000	15.050000	0.184100	0.782300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	50	47.238000	48.180200	47.448800	47.519300	0.344800	0.055300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	100	4.756100	5.322900	5.108600	5.081300	0.192300	0.862700	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	100	4.459200	4.912400	4.803700	4.724500	0.180000	0.309500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	100	12.266300	13.343100	12.378800	12.546700	0.402300	0.007000	No
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	100	37.082000	37.831200	37.661800	37.512800	0.290200	0.351500	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	150	4.511000	5.621500	5.059600	5.085100	0.359300	0.896200	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	150	4.252200	5.106100	4.643900	4.658500	0.272600	0.709000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	150	11.638700	12.215900	11.986600	11.977400	0.211400	0.661300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	150	34.783500	36.672000	35.454000	35.528900	0.633500	0.502800	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	50	200	4.375300	7.077000	5.235600	5.471400	0.932000	0.727700	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	50	200	4.678300	5.142800	4.935500	4.932400	0.155000	0.949400	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	50	200	12.426700	12.895300	12.779400	12.723000	0.168000	0.440700	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	50	200	33.969600	34.973500	34.576400	34.462000	0.385300	0.524300	Yes

Table 44: Descriptive statistics of the latency for RQ6 (Mobile-Cloud Quantized Early Exit Partitioning Models), including Shapiro-Wilk p-values and normality assessment (Part2)

Tier	Model	Operator	Mobile-Edge BW	Edge-Cloud BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Cloud	ResNet	Quantized Early Exit Partition	100	1	12.042500	13.941300	12.427000	12.700600	0.674700	0.266600	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	1	11.598300	12.466600	11.860600	12.026400	0.350300	0.248500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	1	203.737400	204.396800	204.058000	204.102600	0.247200	0.594300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	1	480.772800	481.749800	481.072300	481.142600	0.330900	0.400100	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	10	3.773900	4.276200	4.081100	4.028900	0.183100	0.773900	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	10	27.148600	27.779600	27.687300	27.552100	0.234700	0.261200	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	10	69.464200	72.770800	72.015000	71.667400	1.138800	0.041900	No
Mobile-Cloud	ResNet	Quantized Early Exit Partition	100	50	4.014000	5.013200	4.428000	4.483300	0.346800	0.969900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	50	3.869100	4.316400	4.220500	4.174000	0.156600	0.035300	No
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	50	11.745800	12.122800	12.019700	11.943200	0.144900	0.361000	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	50	33.979000	36.867200	36.570300	36.103100	1.082600	0.014500	No
Mobile-Cloud	ResNet	Quantized Early Exit Partition	100	100	4.046800	4.407900	4.181600	4.212200	0.124700	0.935800	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	100	3.469200	3.806900	3.514800	3.589900	0.127500	0.230500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	100	10.085100	10.603300	10.452900	10.370500	0.222900	0.150400	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	100	32.249300	33.477700	32.401000	32.627900	0.462000	0.114500	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	100	150	3.973600	4.503900	4.388900	4.269900	0.242300	0.053400	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	150	3.545600	4.354000	4.199800	4.048700	0.308500	0.323100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	150	9.819800	11.096600	10.292400	10.348300	0.421000	0.601100	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	150	28.182400	31.573400	30.191900	30.042000	1.083900	0.484000	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	100	200	4.281200	5.177200	4.295400	4.539800	0.350300	0.044200	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	100	200	3.700200	4.362700	3.789900	3.915300	0.252100	0.164000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	100	200	9.968600	10.583200	10.105700	10.172700	0.216700	0.158700	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	100	200	29.746100	31.065000	30.360500	30.407200	0.548700	0.289300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	1	11.949000	12.461400	11.994900	12.076300	0.193300	0.001900	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	1	11.057400	12.322800	11.575900	11.699500	0.446500	0.858500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	1	202.712100	205.165700	203.557800	203.674400	0.809500	0.263300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	1	478.547400	480.351100	479.063100	479.152200	0.631700	0.119400	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	10	3.992400	4.547300	4.322600	4.298800	0.185800	0.945900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	10	3.630000	4.017200	3.790800	3.826400	0.131600	0.908000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	10	26.051200	27.167200	26.799700	26.691400	0.371100	0.743200	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	10	70.353200	72.154600	70.836800	71.107500	0.621400	0.661900	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	50	4.088100	4.346400	4.103800	4.189900	0.113100	0.041700	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	50	3.700100	4.243900	3.915000	3.960800	0.217700	0.410100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	50	11.295100	11.621200	11.509400	11.486300	0.124200	0.537200	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	50	34.536000	35.712900	34.940200	35.106300	0.434100	0.647300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	100	4.203500	4.995000	4.224100	4.446800	0.310400	0.063300	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	100	3.427300	4.443900	4.041000	4.014000	0.341200	0.802000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	100	8.985000	10.125700	9.779300	9.686900	0.385400	0.459300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	100	27.784400	31.417500	29.482600	29.731000	1.465700	0.352500	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	150	3.987500	4.999200	4.361500	4.364300	0.364300	0.405400	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	150	3.541500	4.240600	3.936000	3.932000	0.232700	0.885500	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	150	9.327000	9.842900	9.421700	9.515100	0.196600	0.295900	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	150	26.963300	29.474900	28.872400	28.620200	0.886400	0.220600	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	150	200	3.857100	4.399700	4.214300	4.152300	0.190400	0.857500	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	150	200	3.645000	4.177400	4.135600	3.998200	0.202600	0.130300	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	150	200	9.151400	9.747000	9.539100	9.510700	0.206000	0.662300	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	150	200	26.985500	29.177300	29.029200	28.460300	0.872000	0.100300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	1	11.860500	14.516200	12.409300	12.795600	0.912000	0.167700	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	1	11.144100	11.813000	11.497600	11.445800	0.254200	0.539100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	1	202.773100	203.394600	202.994100	203.047200	0.245500	0.441000	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	1	476.675100	479.823300	478.255600	478.282200	1.010200	0.849000	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	10	4.600500	7.278900	4.770300	5.225300	1.028800	0.000900	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	10	3.806500	5.563900	4.277700	4.421000	0.597300	0.078100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	10	26.963600	28.346700	27.339500	27.526100	0.467800	0.582000	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	10	67.180200	69.849000	69.227100	68.839400	0.911200	0.367700	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	50	4.329800	4.777500	4.583200	4.586000	0.155900	0.823900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	50	4.052300	5.551300	4.380300	4.589400	0.510200	0.165100	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	50	11.681300	12.295600	11.986000	11.987900	0.199400	0.960500	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	50	33.971900	35.208100	34.665600	34.648800	0.427900	0.952900	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	100	4.594800	4.917300	4.722100	4.735500	0.103400	0.306900	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	100	4.171300	4.669600	4.399100	4.422100	0.182300	0.841000	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	100	9.886000	10.404100	10.231500	10.150600	0.207300	0.325000	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	100	29.745200	31.048000	30.442200	30.444800	0.478600	0.811400	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	150	4.323000	5.159000	4.560400	4.670500	0.298400	0.688000	Yes
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	150	4.169500	4.900700	4.258600	4.374500	0.271600	0.034400	No
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	150	9.605000	11.294400	10.091200	10.275100	0.558000	0.324700	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	150	26.695100	29.177400	28.331900	28.028400	0.949300	0.555300	Yes
Mobile-Cloud	ResNet	Quantized Early Exit Partition	200	200	4.584000	6.469200	4.835100	5.084300	0.704300	0.012400	No
Mobile-Cloud	ResNext	Quantized Early Exit Partition	200	200	4.159000	5.115700	4.440400	4.515800	0.319600	0.178700	Yes
Mobile-Cloud	FCN	Quantized Early Exit Partition	200	200	9.558000	10.395800	9.871900	9.943200	0.276900	0.896200	Yes
Mobile-Cloud	DUC	Quantized Early Exit Partition	200	200	25.287100	28.347200	27.240200	26.933600	1.178500	0.555600	Yes

Table 45: Descriptive statistics of the latency for RQ6 (Edge-Cloud Quantized Early Exit Partitioning Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Edge-Cloud BW	Mobile-Edge BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	1	176.072300	178.099300	176.186800	176.566500	0.772200	0.003500	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	1	175.789300	176.090700	175.910300	175.931800	0.105800	0.942000	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	1	212.782700	213.123600	212.914500	212.961300	0.123000	0.625900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	1	528.544200	529.855700	529.665400	529.483500	0.478400	0.020200	No
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	10	168.282300	168.614800	168.496800	168.479300	0.123200	0.615400	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	10	168.017800	168.524200	168.109700	168.189700	0.183100	0.236800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	10	35.587000	36.077100	35.794100	35.789300	0.164000	0.593900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	10	117.951900	118.866100	118.483400	118.457800	0.312100	0.951100	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	50	168.287000	168.500600	168.315200	168.374700	0.093300	0.067500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	50	168.042400	168.375800	168.116500	168.157500	0.114900	0.152900	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	50	19.790600	20.092200	19.929400	19.934300	0.100100	0.993600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	50	81.552700	82.481000	81.812400	81.908300	0.311100	0.350900	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	100	168.384900	170.201800	168.523900	168.821600	0.692800	0.001900	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	100	167.941800	168.269100	168.028700	168.084800	0.122200	0.528100	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	100	17.576000	18.003200	17.786800	17.829700	0.159500	0.346600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	100	76.197600	77.173400	76.852400	76.718700	0.437800	0.094300	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	150	168.285900	169.725800	168.409700	168.639500	0.547500	0.003900	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	150	168.000300	168.078600	168.074700	168.052200	0.031500	0.087700	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	150	17.395400	17.761700	17.429200	17.520300	0.142700	0.141600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	150	74.912600	76.034000	75.542500	75.533900	0.435100	0.534400	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	1	200	168.273800	169.084100	168.382700	168.507300	0.299700	0.049200	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	1	200	168.046500	168.273300	168.169200	168.170700	0.078700	0.937800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	1	200	17.313600	17.702100	17.596300	17.533200	0.166600	0.150700	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	1	200	74.160100	75.337200	74.541400	74.588700	0.401700	0.240100	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	1	24.394300	25.138700	24.812000	24.727500	0.289100	0.385300	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	1	24.267000	24.624300	24.477100	24.470500	0.127700	0.817300	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	1	198.097600	198.183100	198.129900	198.139500	0.035000	0.307400	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	10	1	467.647800	469.172400	468.407400	468.359800	0.533200	0.939700	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	10	16.680400	16.898700	16.812900	16.790400	0.077100	0.864800	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	10	16.365300	16.567000	16.518100	16.498700	0.069500	0.093300	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	10	21.799200	22.260600	21.968700	22.004300	0.165100	0.850900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	10	10	62.368200	62.946200	62.597800	62.607000	0.199600	0.760900	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	50	16.597800	16.865800	16.712400	16.735300	0.089800	0.893400	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	50	16.363200	16.712300	16.535100	16.540400	0.111200	0.704700	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	50	6.101400	6.425600	6.209300	6.253600	0.111200	0.809500	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	10	50	25.878200	26.440300	26.225400	26.158900	0.208500	0.663700	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	100	16.469100	16.971600	16.837500	16.752400	0.188100	0.523900	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	100	16.332500	16.768600	16.647000	16.601000	0.164300	0.456400	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	100	4.032500	4.437500	4.327800	4.294100	0.137300	0.105200	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	10	100	20.558300	21.050600	20.886200	20.855900	0.177400	0.632100	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	150	16.659800	17.207100	16.707700	16.810800	0.203900	0.025500	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	150	16.422900	16.597800	16.516100	16.520500	0.064400	0.719800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	150	3.509500	3.979300	3.940300	3.844500	0.174000	0.037700	No
Edge-Cloud	DUC	Quantized Early Exit Partition	10	150	19.345700	19.836000	19.795700	19.644000	0.210800	0.060900	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	10	200	16.612200	17.030100	16.808500	16.812200	0.135200	0.917000	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	10	200	16.412600	16.656100	16.564400	16.560600	0.082700	0.508600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	10	200	3.901100	4.527300	4.100400	4.207900	0.261200	0.196900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	10	200	18.581600	18.996600	18.873100	18.840900	0.155400	0.436200	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	1	11.177800	11.673400	11.358200	11.385000	0.160400	0.305400	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	1	10.876200	11.548200	11.118300	11.182200	0.251000	0.690600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	1	197.867300	198.113000	198.028100	198.017600	0.082900	0.467600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	50	1	463.040700	465.053200	464.569000	464.332900	0.687000	0.211900	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	10	3.400600	3.630900	3.480500	3.510800	0.082100	0.783000	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	10	3.280200	3.371900	3.294800	3.319900	0.039300	0.087600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	10	21.419500	21.715600	21.541000	21.543100	0.107700	0.691900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	50	10	56.471800	57.188300	56.863800	56.854000	0.268100	0.742200	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	50	3.364700	3.642000	3.450900	3.467200	0.095400	0.381800	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	50	3.177500	3.444700	3.269500	3.304400	0.110300	0.264700	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	50	6.343800	8.322200	6.456800	6.827000	0.752900	0.003300	No
Edge-Cloud	DUC	Quantized Early Exit Partition	50	50	21.542400	22.332400	21.965600	21.900800	0.275400	0.777300	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	100	3.316500	3.537700	3.387500	3.390600	0.080700	0.172300	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	100	2.998500	3.152800	3.040500	3.064000	0.064700	0.196600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	100	3.993000	4.500900	4.344100	4.299600	0.191100	0.515100	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	50	100	16.762400	17.473300	17.150800	17.175400	0.241500	0.652900	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	150	3.318600	3.425300	3.340700	3.352400	0.038600	0.130500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	150	3.057500	3.219300	3.120700	3.124200	0.053300	0.527400	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	150	3.546900	3.871300	3.787500	3.758100	0.110200	0.060000	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	50	150	15.052000	15.652800	15.416000	15.410100	0.205300	0.651600	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	50	200	3.168800	3.459900	3.185800	3.270700	0.116900	0.096800	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	50	200	3.212800	3.405300	3.349500	3.319800	0.070600	0.633500	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	50	200	4.055800	4.167900	4.133700	4.119500	0.037500	0.708700	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	50	200	14.842800	15.264800	15.183000	15.081000	0.193800	0.045600	No
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	1	9.971400	10.344600	10.103100	10.124700	0.128100	0.772700	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	1	9.581700	9.932400	9.658200	9.730700	0.138100	0.295000	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	1	197.993500	198.264800	198.110200	198.104700	0.101100	0.542200	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	1	464.059400	464.793600	464.371600	464.410500	0.235300	0.731500	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	10	2.057400	2.486100	2.123100	2.184200	0.153800	0.020300	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	10	1.818100	2.118000	1.926100	1.938100	0.098400	0.273300	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	10	21.303500	21.661700	21.602500	21.510200	0.141100	0.209500	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	10	56.414200	57.161300	56.777400	56.781200	0.236300	0.339800	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	50	1.983000	2.243500	2.108400	2.118000	0.087300	0.983200	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	50	1.843200	1.966200	1.854500	1.881600	0.045200	0.086500	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	50	5.925600	6.226100	6.000700	6.027900	0.106200	0.220900	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	50	21.069400	21.760000	21.481800	21.442700	0.229300	0.956000	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	100	2.028100	2.121300	2.096900	2.088200	0.031500	0.056900	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	100	1.736500	1.872100	1.820600	1.814100	0.043700	0.318100	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	100	3.978600	4.450200	4.363100	4.301400	0.166700	0.061300	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	100	16.872000	17.779200	17.259700	17.359300	0.350500	0.449400	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	150	2.138300	2.257400	2.167000	2.186500	0.040900	0.532700	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	150	1.803800	1.963900	1.884500	1.883800	0.056800	0.959900	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	150	3.706200	4.402900	3.979000	4.000300	0.225900	0.422600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	150	15.663500	16.546800	16.038200	16.098200	0.319200	0.868400	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	100	200	1.998000	2.252100	2.065400	2.100900	0.089300	0.640300	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	100	200	1.791000	1.946000	1.902000	1.872000	0.064300	0.192900	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	100	200	3.761100	4.310700	4.022900	4.030600	0.197800	0.920000	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	100	200	14.732800	15.333800	15.214300	15.132900	0.209200	0.114100	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	1	9.943800	10.413600	10.009800	10.087500	0.171800	0.089900	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	1	9.730800	10.027600	9.889000	9.901400	0.100600	0.678400	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	150	1	198.014500	198.384000	198.166700	198.166500	0.125000	0.668100	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	1	463.546700	465.223400	464.523500	464.445600	0.534900	0.552500	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	10	1.874200	2.210600	2.119200	2.091800	0.117600	0.292500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	10	1.819300	1.942800	1.844500	1.858900	0.043100	0.041000	No
Edge-Cloud	FCN	Quantized Early Exit Partition	150	10	21.179600	21.484500	21.305300	21.304500	0.112800	0.589800	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	10	55.868100	57.078100	56.550500	56.497600	0.389800	0.850200	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	50	2.022900	2.165600	2.118700	2.097300	0.053300	0.581500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	50	1.830900	2.007700	1.877600	1.915200	0.073800	0.160900	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	150	50	5.922600	6.684800	6.120100	6.204900	0.260700	0.361500	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	50	20.865800	22.047000	21.278900	21.343100	0.385100	0.201000	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	100	1.993400	2.267000	2.112000	2.110200	0.090000	0.631700	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	100	1.794200	2.149700	1.882600	1.925200	0.126900	0.446300	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	150	100	4.189700	4.594500	4.353100	4.370900	0.163600	0.371100	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	100	16.382000	17.384700	16.547600	16.721300	0.363300	0.219700	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	150	2.008800	2.340900	2.163700	2.181500	0.118300	0.899200	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	150	1.754100	2.028800	1.940100	1.898000	0.097200	0.722200	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	150	150	3.723500	4.301000	3.887500	3.942200	0.192600	0.237600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	150	14.868000	15.950900	15.443600	15.379200	0.382100	0.893000	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	150	200	2.096000	2.112800	2.101300	2.103300	0.007000	0.229500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	150	200	1.863300	2.006200	1.958400	1.945400	0.052000	0.729800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	150	200	3.731300	4.101800	3.850900	3.885900	0.138900	0.598700	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	150	200	14.630600	15.349800	15.081600	15.014800	0.276600	0.572400	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	1	9.849700	10.213300	9.938900	9.988700	0.124500	0.436800	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	1	9.723600	9.925800	9.778500	9.808900	0.073500	0.626600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	1	197.999000	198.491000	198.073100	198.158400	0.176100	0.124000	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	200	1	463.179800	464.375800	463.939100	463.894700	0.403400	0.630400	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	10	2.220700	3.818900	2.329600	2.611900	0.606200	0.002200	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	10	1.847700	2.186600	1.993100	2.011500	0.138300	0.352600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	10	21.392500	21.750500	21.633500	21.599800	0.129400	0.734500	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	200	10	56.170100	56.959200	56.678800	56.599700	0.324900	0.292700	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	50	2.176200	2.442500	2.267400	2.284000	0.087500	0.426500	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	50	1.881900	2.170900	1.992600	2.035000	0.109600	0.444600	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	50	6.204000	6.486700	6.373400	6.370700	0.094600	0.620600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	200	50	20.751800	21.791000	21.061000	21.163200	0.343700	0.360300	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	100	2.167200	2.483400	2.299100	2.337400	0.122100	0.426000	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	100	1.906900	2.196700	2.091100	2.086800	0.100600	0.447700	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	100	4.372200	5.156100	4.518000	4.592800	0.290100	0.027500	No
Edge-Cloud	DUC	Quantized Early Exit Partition	200	100	16.774100	17.964600	17.422900	17.443000	0.426900	0.803100	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	150	2.159300	2.645500	2.281300	2.353200	0.166700	0.543100	Yes
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	150	2.052500	2.168900	2.062800	2.092800	0.045400	0.140800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	150	4.059300	4.417400	4.115200	4.181800	0.133000	0.233600	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	200	150	15.030900	15.901100	15.354900	15.459500	0.324300	0.653600	Yes
Edge-Cloud	ResNet	Quantized Early Exit Partition	200	200	2.254300	2.913100	2.273400	2.407000	0.254000	0.001400	No
Edge-Cloud	ResNext	Quantized Early Exit Partition	200	200	1.896000	2.174900	2.066600	2.064400	0.093300	0.423800	Yes
Edge-Cloud	FCN	Quantized Early Exit Partition	200	200	4.041400	4.240000	4.127400	4.136200	0.067600	0.992200	Yes
Edge-Cloud	DUC	Quantized Early Exit Partition	200	200	14.292100	15.171900	14.844300	14.762000	0.311300	0.890400	Yes

Table 46: Descriptive statistics of the latency for RQ6 (Mobile Identity, Quantized, Early Exit, and Quantized Early Exit Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Mobile BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile	DUC	Identity	1	71.115000	74.498700	73.010600	72.844300	1.120600	0.984300	Yes
Mobile	FCN	Identity	1	28.255700	29.457500	28.442400	28.684100	0.433500	0.261700	Yes
Mobile	ResNext	Identity	1	87.602500	91.638800	87.776300	88.521100	1.561900	0.000900	No
Mobile	ResNet	Identity	1	89.913500	90.772700	90.368500	90.358200	0.285300	0.996600	Yes
Mobile	FCN	Identity	10	22.002100	23.293200	22.143100	22.428100	0.471300	0.135300	Yes
Mobile	DUC	Identity	10	44.163700	46.815700	45.393100	45.293500	0.963000	0.671900	Yes
Mobile	ResNet	Identity	10	14.399700	19.240200	14.958800	16.417200	2.158600	0.045500	No
Mobile	ResNext	Identity	10	12.193200	12.621600	12.331600	12.410500	0.167800	0.316500	Yes
Mobile	ResNet	Identity	50	7.872700	8.892800	8.323700	8.384400	0.331600	0.882000	Yes
Mobile	ResNext	Identity	50	5.620900	9.706100	6.016400	6.951100	1.594000	0.128800	Yes
Mobile	DUC	Identity	50	42.826200	43.954900	43.565400	43.404500	0.482800	0.167500	Yes
Mobile	FCN	Identity	50	20.496000	22.995200	22.480900	22.131200	0.857900	0.115100	Yes
Mobile	ResNext	Identity	100	4.930600	7.408600	5.292100	5.902100	0.968700	0.218600	Yes
Mobile	ResNet	Identity	100	7.236700	13.630400	8.056200	9.398600	2.457600	0.181300	Yes
Mobile	DUC	Identity	100	42.065000	42.907700	42.628300	42.582000	0.308600	0.542800	Yes
Mobile	FCN	Identity	100	21.986300	23.592700	22.690200	22.747800	0.534500	0.990200	Yes
Mobile	ResNet	Identity	150	7.683400	13.099500	8.200700	9.360500	2.068900	0.107000	Yes
Mobile	ResNext	Identity	150	5.007000	6.130000	5.409700	5.468800	0.408500	0.692700	Yes
Mobile	DUC	Identity	150	42.850000	44.303600	43.475400	43.510100	0.463200	0.376900	Yes
Mobile	FCN	Identity	150	20.984100	22.972900	22.608700	22.165300	0.790100	0.238700	Yes
Mobile	DUC	Identity	200	43.088100	44.017500	43.803300	43.643000	0.358900	0.373600	Yes
Mobile	ResNext	Identity	200	4.534800	6.230800	5.183300	5.202800	0.602500	0.595300	Yes
Mobile	FCN	Identity	200	21.599400	23.593000	22.489300	22.509200	0.721300	0.859400	Yes
Mobile	ResNet	Identity	200	7.857200	13.293600	9.801700	10.429600	2.097400	0.529900	Yes
Mobile	DUC	Quantization	1	65.947300	70.467900	67.639800	67.962400	1.457800	0.602400	Yes
Mobile	ResNet	Quantization	1	92.306200	100.395100	96.315100	96.620500	2.681800	0.926300	Yes
Mobile	FCN	Quantization	1	21.284200	24.089100	22.690700	22.585400	1.107300	0.463100	Yes
Mobile	ResNext	Quantization	1	88.486600	92.605700	90.079700	90.141100	1.394200	0.599500	Yes
Mobile	ResNet	Quantization	10	13.793900	24.414900	16.530900	17.586500	3.624500	0.208300	Yes
Mobile	DUC	Quantization	10	38.680400	41.244000	40.444000	40.250100	0.851200	0.318900	Yes
Mobile	ResNext	Quantization	10	13.710500	17.391100	14.305600	14.806000	1.347000	0.065500	Yes
Mobile	FCN	Quantization	10	13.811700	18.262300	15.010100	15.617300	1.592600	0.634200	Yes
Mobile	ResNext	Quantization	50	7.416300	11.190700	9.283500	9.164900	1.252900	0.928700	Yes
Mobile	FCN	Quantization	50	15.185200	15.792800	15.397500	15.410700	0.225200	0.399700	Yes
Mobile	DUC	Quantization	50	36.575700	38.422500	38.021600	37.789000	0.660600	0.266700	Yes
Mobile	ResNet	Quantization	50	7.711600	14.729400	8.682300	9.928000	2.510500	0.081500	Yes
Mobile	FCN	Quantization	100	13.675200	15.635900	14.530600	14.727100	0.705600	0.710700	Yes
Mobile	ResNext	Quantization	100	6.137000	9.684600	7.309100	7.504700	1.320500	0.497900	Yes
Mobile	DUC	Quantization	100	37.465400	38.488800	37.963100	37.938300	0.410600	0.402400	Yes
Mobile	ResNet	Quantization	100	7.673700	14.405600	9.305800	10.203200	2.539900	0.401200	Yes
Mobile	ResNext	Quantization	150	6.107700	9.894700	6.909800	7.626000	1.439300	0.383400	Yes
Mobile	FCN	Quantization	150	16.173900	16.930700	16.659300	16.602000	0.255400	0.881900	Yes
Mobile	DUC	Quantization	150	38.105000	39.038400	38.190500	38.423800	0.353700	0.151100	Yes
Mobile	ResNet	Quantization	150	7.922800	10.397700	8.989200	9.263300	0.891500	0.650200	Yes
Mobile	DUC	Quantization	200	37.184800	39.488900	38.389200	38.239900	0.916000	0.399600	Yes
Mobile	FCN	Quantization	200	14.900100	15.589900	15.384000	15.294700	0.234000	0.741000	Yes
Mobile	ResNet	Quantization	200	7.892600	11.711400	10.505400	10.137200	1.320700	0.761100	Yes
Mobile	ResNext	Quantization	200	6.401400	11.006300	7.573100	8.123400	1.554100	0.287800	Yes
Mobile	DUC	Early Exit	1	68.395500	69.322900	69.013700	68.873300	0.351900	0.519000	Yes
Mobile	ResNet	Early Exit	1	89.283200	90.316200	89.679800	89.757300	0.368300	0.888900	Yes
Mobile	FCN	Early Exit	1	25.204800	26.212700	25.715200	25.671900	0.338600	0.929200	Yes
Mobile	ResNext	Early Exit	1	86.795700	87.664700	87.302700	87.202400	0.305300	0.716600	Yes
Mobile	ResNet	Early Exit	10	13.613300	16.128100	14.977400	14.906600	0.860300	0.996400	Yes
Mobile	DUC	Early Exit	10	41.431500	42.313900	41.650000	41.795900	0.354700	0.284000	Yes
Mobile	ResNext	Early Exit	10	11.291900	12.102300	11.725700	11.747300	0.318000	0.446200	Yes
Mobile	FCN	Early Exit	10	18.397500	19.713100	19.198800	19.153800	0.430400	0.642700	Yes
Mobile	ResNext	Early Exit	50	4.790500	5.356400	5.093600	5.045000	0.202400	0.706500	Yes
Mobile	FCN	Early Exit	50	18.771800	20.483700	18.844600	19.162400	0.662400	0.001100	No
Mobile	DUC	Early Exit	50	38.768800	39.765600	39.110100	39.189800	0.339600	0.828600	Yes
Mobile	ResNet	Early Exit	50	7.180400	10.522200	8.021800	8.389700	1.127800	0.152000	Yes
Mobile	FCN	Early Exit	100	19.113800	20.489400	19.894400	19.773200	0.486300	0.846000	Yes
Mobile	ResNext	Early Exit	100	4.720000	7.391600	4.952700	5.417500	0.998000	0.007400	No
Mobile	DUC	Early Exit	100	37.735600	38.909200	38.365100	38.355100	0.377800	0.841300	Yes
Mobile	ResNet	Early Exit	100	6.067300	6.934600	6.392200	6.428600	0.281900	0.392500	Yes
Mobile	ResNext	Early Exit	150	4.410900	4.757500	4.432800	4.523900	0.137700	0.096900	Yes
Mobile	FCN	Early Exit	150	18.908200	19.695200	19.385200	19.309400	0.282000	0.848800	Yes
Mobile	DUC	Early Exit	150	36.997400	40.283600	38.498200	38.669900	1.111500	0.981900	Yes
Mobile	ResNet	Early Exit	150	6.175800	7.086500	6.526700	6.539400	0.327000	0.625600	Yes
Mobile	DUC	Early Exit	200	38.546900	39.799700	39.246100	39.230300	0.459200	0.807800	Yes
Mobile	FCN	Early Exit	200	16.442900	18.394100	18.232700	17.870900	0.723200	0.008800	No
Mobile	ResNet	Early Exit	200	5.775600	7.174100	6.226100	6.419600	0.495900	0.776400	Yes
Mobile	ResNext	Early Exit	200	4.285300	6.235400	4.326200	4.729200	0.756600	0.001400	No
Mobile	DUC	Quantized Early Exit	1	61.508900	64.918500	62.341700	62.725100	1.285200	0.334100	Yes
Mobile	ResNet	Quantized Early Exit	1	88.306300	93.709800	90.410200	91.062400	2.257100	0.226400	Yes
Mobile	FCN	Quantized Early Exit	1	18.593700	21.604800	19.576700	19.849500	0.986200	0.511700	Yes
Mobile	ResNext	Quantized Early Exit	1	88.402900	90.406300	88.514200	88.983600	0.754000	0.046600	No
Mobile	ResNet	Quantized Early Exit	10	14.007700	19.208300	15.813000	15.947900	1.841500	0.444300	Yes
Mobile	DUC	Quantized Early Exit	10	35.555000	38.220100	36.476100	36.500400	0.952600	0.272100	Yes
Mobile	ResNext	Quantized Early Exit	10	13.396400	16.106700	14.511500	14.483700	0.923400	0.554200	Yes
Mobile	FCN	Quantized Early Exit	10	12.250800	14.675700	12.831200	13.139300	0.875100	0.426000	Yes
Mobile	ResNext	Quantized Early Exit	50	6.885500	7.085600	7.015400	6.986400	0.070600	0.726600	Yes
Mobile	FCN	Quantized Early Exit	50	12.994100	14.173200	13.988700	13.732300	0.434200	0.301700	Yes
Mobile	DUC	Quantized Early Exit	50	32.126800	34.512600	34.308400	33.863200	0.881100	0.010900	No
Mobile	ResNet	Quantized Early Exit	50	7.111400	16.908700	11.803400	11.425500	3.345200	0.833800	Yes
Mobile	FCN	Quantized Early Exit	100	13.007800	13.987100	13.633500	13.594600	0.337000	0.683400	Yes
Mobile	ResNext	Quantized Early Exit	100	5.719800	7.696600	7.013700	6.812900	0.704900	0.834300	Yes
Mobile	DUC	Quantized Early Exit	100	31.821400	35.566700	34.162000	33.978200	1.212800	0.441600	Yes
Mobile	ResNet	Quantized Early Exit	100	6.801600	11.592200	8.314400	9.005100	1.947200	0.300100	Yes
Mobile	ResNext	Quantized Early Exit	150	5.431600	6.884800	6.105300	6.106300	0.487700	0.981900	Yes
Mobile	FCN	Quantized Early Exit	150	11.588000	13.100300	12.381200	12.388700	0.591300	0.533900	Yes
Mobile	DUC	Quantized Early Exit	150	30.422400	34.293400	32.947700	32.870400	1.338700	0.318500	Yes
Mobile	ResNet	Quantized Early Exit	150	6.610700	11.712300	9.299400	8.905600	1.762200	0.773400	Yes
Mobile	DUC	Quantized Early Exit	200	30.364200	34.808300	33.553900	33.165400	1.625000	0.451800	Yes
Mobile	FCN	Quantized Early Exit	200	12.861600	14.637200	13.999300	13.724300	0.654000	0.546400	Yes
Mobile	ResNet	Quantized Early Exit	200	6.714000	9.509900	7.406800	7.744900	0.961600	0.363200	Yes
Mobile	ResNext	Quantized Early Exit	200	5.099100	6.003500	5.613200	5.546800	0.346000	0.609500	Yes

Table 47: Descriptive statistics of the latency for RQ6 (Edge Identity, Quantized, Early Exit, and Quantized Early Exit Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Mobile BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Edge	DUC	Identity	1	80.387600	81.998600	81.988100	81.487500	0.656700	0.052400	Yes
Edge	ResNet	Identity	1	171.571600	174.023900	171.653300	172.105700	0.960200	0.000600	No
Edge	FCN	Identity	1	23.624200	24.588000	24.306000	24.256400	0.347500	0.262000	Yes
Edge	ResNext	Identity	1	169.594000	170.237400	169.697500	169.788000	0.238100	0.092600	Yes
Edge	ResNet	Identity	10	19.795400	20.116500	20.020600	19.983200	0.113800	0.727000	Yes
Edge	DUC	Identity	10	24.502600	25.953000	24.852500	25.179600	0.576700	0.243000	Yes
Edge	ResNext	Identity	10	18.092700	19.631900	18.243300	18.655300	0.617000	0.134300	Yes
Edge	FCN	Identity	10	10.303200	10.899200	10.510400	10.562700	0.201900	0.875900	Yes
Edge	ResNext	Identity	50	4.765300	5.320300	4.900400	4.944100	0.199700	0.137100	Yes
Edge	FCN	Identity	50	9.873000	10.905900	10.594000	10.406100	0.405400	0.343700	Yes
Edge	DUC	Identity	50	19.047700	20.562100	20.095200	20.014000	0.531400	0.364800	Yes
Edge	ResNet	Identity	50	6.801600	7.909800	6.978800	7.164000	0.393700	0.125000	Yes
Edge	FCN	Identity	100	9.331000	10.785500	10.294000	10.176900	0.555400	0.537400	Yes
Edge	ResNext	Identity	100	3.420200	4.103700	3.705400	3.804400	0.260000	0.337000	Yes
Edge	DUC	Identity	100	19.166500	23.449800	22.998200	22.100100	1.593000	0.114800	Yes
Edge	ResNet	Identity	100	5.186000	8.729800	7.196400	6.829000	1.302300	0.641000	Yes
Edge	ResNext	Identity	150	3.392000	5.625500	3.974500	4.135300	0.775800	0.080900	Yes
Edge	FCN	Identity	150	10.094400	10.863600	10.690900	10.606000	0.269900	0.160000	Yes
Edge	DUC	Identity	150	19.110600	23.287900	21.924700	21.415900	1.687200	0.337500	Yes
Edge	ResNet	Identity	150	5.166400	7.321900	5.299400	5.721800	0.814900	0.010500	No
Edge	DUC	Identity	200	18.175100	23.084000	19.987100	20.170600	1.781800	0.669000	Yes
Edge	FCN	Identity	200	10.465500	10.943800	10.778200	10.715400	0.168800	0.804400	Yes
Edge	ResNet	Identity	200	5.004700	5.817500	5.515500	5.437400	0.298600	0.766800	Yes
Edge	ResNext	Identity	200	3.502600	4.720400	3.743100	4.040000	0.514500	0.123700	Yes
Edge	DUC	Quantization	1	79.005600	80.234100	79.686500	79.710800	0.460600	0.595500	Yes
Edge	FCN	Quantization	1	18.748300	19.403300	18.921900	19.044900	0.248900	0.441800	Yes
Edge	ResNext	Quantization	1	169.004500	172.702300	169.090600	169.862800	1.432100	0.002600	No
Edge	ResNet	Quantization	1	169.599500	175.601200	169.909000	171.122400	2.270400	0.006900	No
Edge	FCN	Quantization	10	5.578500	6.089600	5.586600	5.690000	0.200300	0.000800	No
Edge	DUC	Quantization	10	21.886600	23.332400	22.759800	22.671100	0.482400	0.955600	Yes
Edge	ResNet	Quantization	10	18.377600	25.114000	20.404100	21.329900	2.831300	0.207300	Yes
Edge	ResNext	Quantization	10	17.300000	20.303200	17.412800	18.009700	1.151900	0.001800	No
Edge	ResNet	Quantization	50	4.886800	11.806100	11.006300	9.521800	2.637500	0.160000	Yes
Edge	ResNext	Quantization	50	4.307100	7.190400	6.018300	5.726600	1.215600	0.217600	Yes
Edge	DUC	Quantization	50	18.783600	19.283100	19.194600	19.090500	0.188500	0.327600	Yes
Edge	FCN	Quantization	50	5.370700	6.227000	5.488100	5.669400	0.321500	0.223600	Yes
Edge	ResNext	Quantization	100	2.695600	5.410300	2.709300	3.307900	1.058400	0.001600	No
Edge	ResNet	Quantization	100	3.408500	7.496400	3.908900	4.788500	1.597000	0.163200	Yes
Edge	DUC	Quantization	100	18.371800	20.276800	19.486800	19.388400	0.667800	0.954500	Yes
Edge	FCN	Quantization	100	5.170900	6.212800	5.482700	5.649900	0.392600	0.535900	Yes
Edge	ResNet	Quantization	150	3.401100	9.989900	3.517200	5.867000	2.947400	0.025900	No
Edge	ResNext	Quantization	150	2.767300	4.601500	2.808700	3.297600	0.707600	0.043900	No
Edge	DUC	Quantization	150	18.187000	19.315000	18.989200	18.806800	0.410700	0.637300	Yes
Edge	FCN	Quantization	150	5.419700	6.365600	5.521700	5.772200	0.378200	0.146600	Yes
Edge	DUC	Quantization	200	17.484700	20.007100	19.022800	18.902000	0.820600	0.722600	Yes
Edge	ResNext	Quantization	200	2.801600	4.813600	3.048800	3.374200	0.741100	0.034600	No
Edge	FCN	Quantization	200	5.361800	6.905700	5.774500	5.865100	0.545000	0.093500	Yes
Edge	ResNet	Quantization	200	3.207000	6.714600	3.764900	4.538800	1.353500	0.259400	Yes
Edge	DUC	Early Exit	1	77.682500	81.893500	81.324800	80.115000	1.943200	0.033000	No
Edge	FCN	Early Exit	1	22.221400	22.626300	22.443600	22.455300	0.144400	0.710700	Yes
Edge	ResNext	Early Exit	1	169.225900	169.691200	169.392800	169.414900	0.152100	0.336700	Yes
Edge	ResNet	Early Exit	1	170.401900	171.446900	170.590900	170.697800	0.382300	0.018900	No
Edge	FCN	Early Exit	10	8.013100	9.781600	9.064000	9.009400	0.571400	0.581000	Yes
Edge	DUC	Early Exit	10	21.072800	23.669800	22.421900	22.303500	0.885900	0.961900	Yes
Edge	ResNet	Early Exit	10	18.865200	19.910900	19.252000	19.303500	0.373900	0.777400	Yes
Edge	ResNext	Early Exit	10	17.623300	18.027900	17.792500	17.828200	0.166500	0.296300	Yes
Edge	ResNet	Early Exit	50	5.398500	6.318400	5.572300	5.694700	0.321300	0.046600	No
Edge	ResNext	Early Exit	50	4.221400	4.603000	4.329900	4.387500	0.139000	0.648800	Yes
Edge	DUC	Early Exit	50	20.453400	21.098900	20.657400	20.695000	0.232600	0.490000	Yes
Edge	FCN	Early Exit	50	8.972400	9.632500	9.465700	9.376400	0.261000	0.314000	Yes
Edge	ResNext	Early Exit	100	2.979200	3.693300	3.027900	3.242300	0.293900	0.087700	Yes
Edge	ResNet	Early Exit	100	4.174700	4.798600	4.304900	4.378100	0.225500	0.177000	Yes
Edge	DUC	Early Exit	100	20.329800	21.047600	20.708700	20.659300	0.253800	0.824300	Yes
Edge	FCN	Early Exit	100	8.573700	9.787000	9.456700	9.347300	0.422400	0.368600	Yes
Edge	ResNet	Early Exit	150	4.170800	5.426300	5.021900	4.907300	0.450200	0.728500	Yes
Edge	ResNext	Early Exit	150	2.696300	3.606600	3.283100	3.227100	0.295600	0.379100	Yes
Edge	DUC	Early Exit	150	20.535100	20.924300	20.827700	20.771300	0.142200	0.527800	Yes
Edge	FCN	Early Exit	150	8.381100	9.511600	9.136200	8.992100	0.478100	0.203300	Yes
Edge	DUC	Early Exit	200	17.536200	20.718600	20.261100	19.808300	1.167500	0.033900	No
Edge	ResNext	Early Exit	200	3.011500	4.404700	3.098800	3.540200	0.595000	0.058800	Yes
Edge	FCN	Early Exit	200	8.776200	9.667200	9.224600	9.169900	0.306000	0.776100	Yes
Edge	ResNet	Early Exit	200	4.090200	5.646600	4.554400	4.729400	0.581400	0.574000	Yes
Edge	DUC	Quantized EarlyExit	1	76.907400	78.591100	77.841400	77.901800	0.619500	0.548100	Yes
Edge	ResNet	Quantized EarlyExit	1	169.199900	173.201900	169.329100	170.346800	1.536200	0.043700	No
Edge	FCN	Quantized EarlyExit	1	18.198800	19.016800	18.385900	18.512600	0.295600	0.483800	Yes
Edge	ResNext	Quantized EarlyExit	1	168.696100	171.200700	169.097100	169.680900	1.023300	0.167500	Yes
Edge	ResNet	Quantized EarlyExit	10	17.686500	22.121100	17.897500	18.842200	1.691200	0.016800	No
Edge	DUC	Quantized EarlyExit	10	19.683700	21.502000	20.726700	20.659700	0.636500	0.958700	Yes
Edge	ResNext	Quantized EarlyExit	10	17.011500	17.281100	17.091800	17.124200	0.102400	0.527300	Yes
Edge	FCN	Quantized EarlyExit	10	4.508400	4.991500	4.590700	4.651200	0.172900	0.012500	No
Edge	ResNext	Quantized EarlyExit	50	3.705900	4.129200	3.914800	3.949600	0.152600	0.556200	Yes
Edge	FCN	Quantized EarlyExit	50	4.499900	4.706100	4.670500	4.632100	0.078900	0.266400	Yes
Edge	DUC	Quantized EarlyExit	50	16.879900	17.454700	17.136600	17.149500	0.189500	0.976400	Yes
Edge	ResNet	Quantized EarlyExit	50	4.090600	8.091500	4.201900	4.938000	1.577500	0.000300	No
Edge	FCN	Quantized EarlyExit	100	4.387600	5.212200	4.670500	4.708900	0.289800	0.609900	Yes
Edge	ResNext	Quantized EarlyExit	100	2.406300	3.112200	2.579900	2.660900	0.244800	0.329100	Yes
Edge	DUC	Quantized EarlyExit	100	15.102300	17.607900	16.685700	16.573000	0.814300	0.312900	Yes
Edge	ResNet	Quantized EarlyExit	100	2.803100	5.602400	4.316400	4.123200	1.114000	0.413300	Yes
Edge	ResNext	Quantized EarlyExit	150	2.401000	2.506000	2.479000	2.459900	0.046500	0.094700	Yes
Edge	FCN	Quantized EarlyExit	150	4.608200	4.951600	4.800900	4.782700	0.152000	0.131700	Yes
Edge	DUC	Quantized EarlyExit	150	16.661300	17.174900	17.051600	17.021000	0.188600	0.056500	Yes
Edge	ResNet	Quantized EarlyExit	150	3.071600	5.748900	3.796000	4.022700	0.988600	0.354400	Yes
Edge	DUC	Quantized EarlyExit	200	15.027100	16.362100	15.925500	15.868300	0.450200	0.213000	Yes
Edge	FCN	Quantized EarlyExit	200	4.547800	4.791900	4.639800	4.651700	0.083800	0.865300	Yes
Edge	ResNet	Quantized EarlyExit	200	2.914600	5.528800	3.263200	4.060600	1.151100	0.050700	Yes
Edge	ResNext	Quantized EarlyExit	200	2.337100	2.617100	2.462800	2.478500	0.109700	0.519500	Yes

Table 48: Descriptive statistics of the latency for RQ6 (Mobile-Edge Partition and Mobile-Edge Quantized Early Exit Partition Models), including Shapiro-Wilk p-values and normality assessment

Tier	Model	Operator	Mobile-Edge BW	Min	Max	Median	Mean	Std	Shapiro-Wilk p	Normal?
Mobile-Edge	DUC	Partition	1	1213.411300	1214.121900	1213.797800	1213.794600	0.231100	0.876800	Yes
Mobile-Edge	ResNet	Partition	1	96.984500	97.409300	97.117600	97.161500	0.141100	0.638100	Yes
Mobile-Edge	FCN	Partition	1	220.971700	222.211200	222.044500	221.699500	0.559600	0.045500	No
Mobile-Edge	ResNext	Partition	1	95.031800	96.405500	95.127400	95.382700	0.518500	0.008000	No
Mobile-Edge	ResNet	Partition	10	13.961100	14.395900	14.119000	14.155600	0.166200	0.591400	Yes
Mobile-Edge	DUC	Partition	10	157.121200	158.564800	157.462000	157.745000	0.566500	0.327100	Yes
Mobile-Edge	ResNext	Partition	10	11.621000	11.990700	11.708900	11.766000	0.138100	0.461300	Yes
Mobile-Edge	FCN	Partition	10	37.581600	38.422100	38.300600	38.178600	0.302600	0.009300	No
Mobile-Edge	ResNext	Partition	50	4.934200	6.907800	5.387300	5.664400	0.679200	0.347600	Yes
Mobile-Edge	FCN	Partition	50	22.086400	23.075500	22.141100	22.388100	0.376600	0.073700	Yes
Mobile-Edge	DUC	Partition	50	62.883300	64.117600	64.001700	63.810400	0.468700	0.005700	No
Mobile-Edge	ResNet	Partition	50	7.242300	8.545200	7.917600	7.824700	0.510800	0.420900	Yes
Mobile-Edge	FCN	Partition	100	18.180700	20.791000	20.005000	19.803300	0.913600	0.526100	Yes
Mobile-Edge	ResNext	Partition	100	4.597700	5.323900	4.700000	4.825500	0.271000	0.134900	Yes
Mobile-Edge	DUC	Partition	100	52.297700	53.008600	52.662700	52.663100	0.257000	0.903400	Yes
Mobile-Edge	ResNet	Partition	100	6.637300	7.886800	6.876000	7.158800	0.481800	0.302400	Yes
Mobile-Edge	ResNext	Partition	150	4.403900	5.414900	4.595000	4.739900	0.360400	0.181200	Yes
Mobile-Edge	FCN	Partition	150	18.407300	20.188200	19.364700	19.367900	0.574900	0.785500	Yes
Mobile-Edge	DUC	Partition	150	47.440100	49.001600	48.477200	48.413500	0.530200	0.359700	Yes
Mobile-Edge	ResNet	Partition	150	6.822700	7.887000	7.115400	7.186100	0.381500	0.248600	Yes
Mobile-Edge	DUC	Partition	200	44.827500	47.439200	46.791700	46.540800	0.948600	0.292100	Yes
Mobile-Edge	FCN	Partition	200	19.064900	19.832800	19.367000	19.414800	0.271400	0.916900	Yes
Mobile-Edge	ResNet	Partition	200	6.694100	7.796000	6.805400	7.118600	0.446000	0.124000	Yes
Mobile-Edge	ResNext	Partition	200	4.434300	4.582900	4.500700	4.505500	0.048000	0.897100	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	1	515.998400	516.711100	516.422600	516.412300	0.238500	0.737000	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	1	95.412400	95.708600	95.605400	95.604800	0.106200	0.232600	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	1	213.207400	214.498900	213.797400	213.813300	0.451900	0.967900	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	1	94.412500	95.303000	95.007600	94.963300	0.335600	0.442500	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	10	12.099200	12.377800	12.191700	12.215900	0.099400	0.766900	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	10	74.757100	76.610100	76.463800	75.970200	0.749800	0.107000	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	10	11.437500	11.875400	11.706900	11.669900	0.165100	0.672800	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	10	28.691300	30.174600	29.493100	29.477000	0.519500	0.967500	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	50	5.179400	5.791800	5.512500	5.482500	0.251500	0.327800	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	50	12.989900	13.980700	13.345200	13.476600	0.359200	0.718100	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	50	38.190100	39.405500	38.985700	38.917300	0.403400	0.517200	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	50	5.696700	6.297000	5.968400	5.937000	0.222500	0.444700	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	100	11.302200	12.838300	11.840500	11.916500	0.502900	0.237700	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	100	4.100600	4.808300	4.432600	4.430000	0.258200	0.841700	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	100	31.008100	35.741400	32.629600	33.340000	1.764000	0.575400	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	100	5.116100	6.008000	5.224500	5.407000	0.328500	0.149100	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	150	4.294800	4.791700	4.506700	4.558500	0.174000	0.771200	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	150	10.985400	11.912800	11.470600	11.451800	0.382900	0.319100	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	150	31.959400	33.530600	32.732900	32.790500	0.509200	0.807500	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	150	5.005700	5.796100	5.192800	5.264200	0.275900	0.074700	Yes
Mobile-Edge	DUC	Quantized Early Exit Partition	200	29.136400	31.192500	30.531400	30.227600	0.732900	0.734800	Yes
Mobile-Edge	FCN	Quantized Early Exit Partition	200	10.843200	11.599100	11.306400	11.230700	0.265200	0.893700	Yes
Mobile-Edge	ResNet	Quantized Early Exit Partition	200	5.106600	5.310700	5.203100	5.205800	0.088300	0.180500	Yes
Mobile-Edge	ResNext	Quantized Early Exit Partition	200	4.104500	4.716000	4.509200	4.447700	0.253500	0.261700	Yes

Table 49: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0028, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0008.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.016	0.001	0.001	0.378
50 vs 150	0.378	0.001	0.001	0.774
50 vs 200	0.001	0.001	0.001	0.924
100 vs 150	0.497	0.275	0.924	0.378
100 vs 200	0.631	0.631	0.037	0.275
150 vs 200	0.378	0.189	0.005	0.924

Table 50: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0266, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0010.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.497	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.924
50 vs 150	0.924	0.001	0.001	0.378
50 vs 200	0.631	0.001	0.001	0.774
100 vs 150	0.631	0.497	0.071	0.924
100 vs 200	0.924	0.924	0.122	0.924
150 vs 200	0.497	0.378	0.497	0.497

Table 51: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0008, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.016	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.001	0.122
50 vs 150	0.189	0.001	0.001	0.631
50 vs 200	0.497	0.001	0.001	0.189
100 vs 150	0.497	0.189	0.497	0.378
100 vs 200	0.005	0.189	0.016	0.016
150 vs 200	0.071	0.774	0.924	0.189

Table 52: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0033, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0009.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.275	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.631
50 vs 150	0.189	0.001	0.001	0.924
50 vs 200	0.497	0.001	0.001	0.497
100 vs 150	0.122	0.924	0.037	0.631
100 vs 200	0.631	0.774	0.774	0.774
150 vs 200	0.631	0.774	0.275	0.275

Table 53: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0120, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0005.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.497	0.001	0.001	0.001
10 vs 100	0.378	0.001	0.001	0.001
10 vs 150	0.189	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.122
50 vs 150	0.497	0.001	0.001	0.071
50 vs 200	0.378	0.001	0.001	0.275
100 vs 150	0.774	0.071	0.774	0.774
100 vs 200	0.497	0.037	0.774	0.774
150 vs 200	0.631	0.497	0.631	0.631

Table 54: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0137, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0004.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.631	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.378	0.001	0.001	0.001
10 vs 200	0.497	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.774
50 vs 150	0.189	0.001	0.001	0.631
50 vs 200	0.774	0.001	0.001	0.016
100 vs 150	0.071	0.037	0.189	0.631
100 vs 200	0.497	0.071	0.275	0.122
150 vs 200	0.631	0.378	0.378	0.189

Table 55: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0057, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0008.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.071	0.001	0.001	0.774
50 vs 150	0.189	0.001	0.071	0.189
50 vs 200	0.016	0.001	0.001	0.924
100 vs 150	0.774	0.924	0.071	0.275
100 vs 200	0.924	0.071	0.378	0.631
150 vs 200	0.275	0.001	0.378	0.378

Table 56: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0309, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0011.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.631	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.774	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.924
50 vs 150	0.924	0.001	0.001	0.924
50 vs 200	0.631	0.001	0.001	0.497
100 vs 150	0.924	0.189	0.631	0.631
100 vs 200	0.774	0.037	0.378	0.924
150 vs 200	0.924	0.378	0.497	0.774

Table 57: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0020, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.016	0.001	0.001	0.001
10 vs 200	0.189	0.001	0.001	0.001
50 vs 100	0.071	0.001	0.001	0.122
50 vs 150	0.122	0.001	0.001	0.001
50 vs 200	0.631	0.001	0.001	0.497
100 vs 150	0.774	0.774	0.774	0.924
100 vs 200	0.189	0.497	0.071	0.275
150 vs 200	0.122	0.631	0.497	0.005

Table 58: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0010, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0007.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.037	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.378	0.001	0.001	0.924
50 vs 150	0.189	0.001	0.001	0.378
50 vs 200	0.005	0.001	0.001	0.189
100 vs 150	0.631	0.189	0.497	0.631
100 vs 200	0.016	0.189	0.071	0.275
150 vs 200	0.037	0.631	0.122	0.497

Table 59: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0039, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.189	0.001	0.001	0.001
10 vs 150	0.037	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.275	0.001	0.001	0.122
50 vs 150	0.924	0.001	0.001	0.774
50 vs 200	0.378	0.001	0.001	0.497
100 vs 150	0.275	0.275	0.122	0.122
100 vs 200	0.275	0.275	0.924	0.016
150 vs 200	0.189	0.378	0.037	0.189

Table 60: Pairwise Conover post-hoc p-values for RQ6 (Cloud Identity Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0033, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.071	0.001	0.001	0.001
10 vs 150	0.016	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.189	0.001	0.001	0.001
50 vs 150	0.189	0.001	0.001	0.016
50 vs 200	0.497	0.001	0.001	0.924
100 vs 150	0.378	0.774	0.631	0.924
100 vs 200	0.378	0.189	0.378	0.001
150 vs 200	0.275	0.122	0.774	0.001

Table 61: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0130, ResNet = 0.0002, ResNext = 0.0001, DUC = 0.0004.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.275	0.001	0.001	0.631
50 vs 150	0.774	0.001	0.001	0.005
50 vs 200	0.275	0.122	0.001	0.275
100 vs 150	0.924	0.071	0.497	0.122
100 vs 200	0.122	0.497	0.924	0.774
150 vs 200	0.275	0.497	0.774	0.378

Table 62: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0286, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.924	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.631	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.275
50 vs 150	0.774	0.001	0.001	0.275
50 vs 200	0.378	0.001	0.001	0.005
100 vs 150	0.774	0.774	0.071	0.924
100 vs 200	0.774	0.275	0.275	0.275
150 vs 200	0.924	0.189	0.275	0.037

Table 63: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0018, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.001	0.001	0.001
10 vs 100	0.005	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.189	0.001	0.001	0.001
50 vs 100	0.016	0.001	0.001	0.001
50 vs 150	0.378	0.122	0.001	0.001
50 vs 200	0.774	0.001	0.001	0.071
100 vs 150	0.275	0.071	0.037	0.924
100 vs 200	0.037	0.037	0.275	0.774
150 vs 200	0.378	0.774	0.005	0.774

Table 64: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0012, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.071	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.001	0.016
50 vs 150	0.016	0.001	0.001	0.774
50 vs 200	0.001	0.001	0.001	0.631
100 vs 150	0.497	0.189	0.275	0.122
100 vs 200	0.497	0.497	0.122	0.001
150 vs 200	0.774	0.774	0.774	0.924

Table 65: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0047, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.189	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.005	0.001
50 vs 150	0.378	0.001	0.001	0.497
50 vs 200	0.774	0.001	0.001	0.016
100 vs 150	0.497	0.631	0.924	0.189
100 vs 200	0.497	0.631	0.497	0.275
150 vs 200	0.275	0.924	0.378	0.497

Table 66: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0031, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.037	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.189	0.001	0.001	0.071
50 vs 150	0.774	0.001	0.001	0.001
50 vs 200	0.631	0.001	0.001	0.189
100 vs 150	0.378	0.774	0.924	0.001
100 vs 200	0.275	0.275	0.631	0.001
150 vs 200	0.924	0.037	0.631	0.005

Table 67: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0013, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.378	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.016	0.001	0.001	0.275
50 vs 150	0.001	0.001	0.001	0.497
50 vs 200	0.005	0.001	0.001	0.016
100 vs 150	0.037	0.071	0.275	0.275
100 vs 200	0.497	0.189	0.275	0.071
150 vs 200	0.189	0.122	0.497	0.037

Table 68: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0308, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.924
50 vs 150	0.774	0.001	0.001	0.189
50 vs 200	0.774	0.001	0.001	0.497
100 vs 150	0.774	0.924	0.631	0.497
100 vs 200	0.924	0.275	0.275	0.189
150 vs 200	0.631	0.275	0.071	0.189

Table 69: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0004, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0004.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.005	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.016	0.001	0.001	0.122
50 vs 150	0.016	0.001	0.001	0.497
50 vs 200	0.774	0.001	0.001	0.071
100 vs 150	0.001	0.924	0.631	0.275
100 vs 200	0.016	0.275	0.631	0.631
150 vs 200	0.037	0.016	0.924	0.189

Table 70: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0010, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0009.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.005	0.001	0.001	0.001
10 vs 100	0.005	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.016	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.275
50 vs 150	0.378	0.001	0.001	0.497
50 vs 200	0.071	0.001	0.001	0.631
100 vs 150	0.631	0.497	0.378	0.774
100 vs 200	0.189	0.378	0.037	0.924
150 vs 200	0.497	0.275	0.122	0.631

Table 71: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0029, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.016	0.001	0.001	0.001
10 vs 200	0.016	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.497
50 vs 150	0.631	0.001	0.001	0.497
50 vs 200	0.497	0.001	0.001	0.001
100 vs 150	0.378	0.924	0.774	0.924
100 vs 200	0.631	0.005	0.037	0.001
150 vs 200	0.774	0.037	0.122	0.071

Table 72: Pairwise Conover post-hoc p-values for RQ6 (Cloud Early Exit Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0013, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.275	0.001	0.001	0.122
50 vs 150	0.071	0.001	0.001	0.924
50 vs 200	0.774	0.001	0.001	0.122
100 vs 150	0.497	0.774	0.378	0.275
100 vs 200	0.497	0.037	0.005	0.001
150 vs 200	0.122	0.001	0.001	0.005

Table 73: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0146, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.631	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.631	0.001	0.001	0.001
50 vs 100	0.189	0.001	0.001	0.016
50 vs 150	0.774	0.001	0.001	0.497
50 vs 200	0.189	0.001	0.001	0.071
100 vs 150	0.378	0.378	0.924	0.005
100 vs 200	0.631	0.189	0.378	0.497
150 vs 200	0.378	0.071	0.631	0.016

Table 74: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0313, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.774	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.378
50 vs 150	0.924	0.001	0.016	0.071
50 vs 200	0.774	0.001	0.005	0.005
100 vs 150	0.774	0.631	0.071	0.497
100 vs 200	0.774	0.122	0.275	0.001
150 vs 200	0.924	0.122	0.189	0.001

Table 75: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0016, ResNet = 0.0000, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.005	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.631
50 vs 150	0.631	0.001	0.001	0.924
50 vs 200	0.275	0.001	0.001	0.122
100 vs 150	0.631	0.275	0.378	0.497
100 vs 200	0.497	0.001	0.071	0.005
150 vs 200	0.924	0.001	0.631	0.037

Table 76: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0055, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.378	0.001	0.001	0.001
10 vs 100	0.037	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.497
50 vs 150	0.189	0.001	0.001	0.001
50 vs 200	0.378	0.001	0.001	0.378
100 vs 150	0.631	0.378	0.016	0.001
100 vs 200	0.774	0.189	0.016	0.275
150 vs 200	0.924	0.631	0.378	0.037

Table 77: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0058, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.189	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.189	0.001	0.001	0.001
50 vs 150	0.631	0.001	0.001	0.497
50 vs 200	0.378	0.001	0.001	0.774
100 vs 150	0.378	0.189	0.774	0.016
100 vs 200	0.924	0.497	0.275	0.275
150 vs 200	0.497	0.631	0.189	0.774

Table 78: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0035, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0007.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.037	0.001	0.001	0.001
10 vs 150	0.016	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.275
50 vs 150	0.631	0.001	0.001	0.378
50 vs 200	0.774	0.001	0.001	0.924
100 vs 150	0.924	0.924	0.774	0.378
100 vs 200	0.497	0.378	0.924	0.189
150 vs 200	0.924	0.497	0.497	0.924

Table 79: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0253, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.631	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.378
50 vs 150	0.924	0.001	0.001	0.122
50 vs 200	0.631	0.001	0.001	0.189
100 vs 150	0.631	0.275	0.924	0.774
100 vs 200	0.924	0.924	0.016	0.378
150 vs 200	0.122	0.497	0.189	0.275

Table 80: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0299, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.122
50 vs 150	0.924	0.001	0.001	0.497
50 vs 200	0.924	0.001	0.001	0.071
100 vs 150	0.774	0.631	0.497	0.037
100 vs 200	0.774	0.924	0.189	0.631
150 vs 200	0.774	0.631	0.378	0.005

Table 81: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0024, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.122
50 vs 150	0.016	0.001	0.001	0.774
50 vs 200	0.275	0.001	0.001	0.497
100 vs 150	0.037	0.924	0.378	0.001
100 vs 200	0.189	0.924	0.924	0.037
150 vs 200	0.071	0.774	0.924	0.924

Table 82: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0100, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.001
50 vs 150	0.924	0.001	0.001	0.016
50 vs 200	0.924	0.001	0.001	0.378
100 vs 150	0.924	0.275	0.016	0.924
100 vs 200	0.631	0.071	0.001	0.071
150 vs 200	0.924	0.071	0.275	0.122

Table 83: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0076, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0009.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.774
50 vs 150	0.924	0.001	0.001	0.924
50 vs 200	0.631	0.001	0.001	0.924
100 vs 150	0.924	0.631	0.631	0.378
100 vs 200	0.631	0.005	0.924	0.378
150 vs 200	0.497	0.001	0.037	0.497

Table 84: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0101, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.275
50 vs 150	0.924	0.001	0.001	0.378
50 vs 200	0.631	0.001	0.001	0.001
100 vs 150	0.924	0.774	0.774	0.497
100 vs 200	0.774	0.122	0.497	0.016
150 vs 200	0.924	0.016	0.001	0.774

Table 85: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0222, ResNet = 0.0002, ResNext = 0.0001, DUC = 0.0011.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.378	0.001	0.001	0.001
10 vs 200	0.774	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.924
50 vs 150	0.924	0.001	0.001	0.774
50 vs 200	0.497	0.122	0.001	0.631
100 vs 150	0.774	0.924	0.378	0.774
100 vs 200	0.631	0.189	0.378	0.774
150 vs 200	0.189	0.189	0.071	0.774

Table 86: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0314, ResNet = 0.0001, ResNext = 0.0002, DUC = 0.0002.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.005
50 vs 150	0.774	0.001	0.001	0.037
50 vs 200	0.774	0.001	0.016	0.001
100 vs 150	0.924	0.275	0.497	0.631
100 vs 200	0.924	0.037	0.924	0.275
150 vs 200	0.924	0.497	0.378	0.189

Table 87: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0010, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.005	0.001	0.001	0.001
10 vs 200	0.005	0.001	0.001	0.001
50 vs 100	0.122	0.001	0.001	0.001
50 vs 150	0.378	0.001	0.001	0.001
50 vs 200	0.631	0.001	0.001	0.001
100 vs 150	0.924	0.071	0.924	0.071
100 vs 200	0.275	0.037	0.001	0.189
150 vs 200	0.189	0.774	0.001	0.497

Table 88: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0009, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.275	0.001	0.001	0.001
10 vs 100	0.071	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.001	0.189
50 vs 150	0.016	0.001	0.001	0.001
50 vs 200	0.005	0.001	0.001	0.005
100 vs 150	0.071	0.924	0.631	0.001
100 vs 200	0.924	0.924	0.275	0.037
150 vs 200	0.275	0.774	0.275	0.774

Table 89: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0057, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.122	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.122	0.001	0.001	0.001
50 vs 150	0.189	0.001	0.001	0.497
50 vs 200	0.037	0.001	0.001	0.631
100 vs 150	0.631	0.924	0.378	0.122
100 vs 200	0.631	0.189	0.497	0.122
150 vs 200	0.774	0.071	0.497	0.924

Table 90: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0015, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.037	0.001	0.001	0.001
10 vs 200	0.005	0.001	0.001	0.001
50 vs 100	0.275	0.001	0.001	0.189
50 vs 150	0.497	0.122	0.001	0.001
50 vs 200	0.122	0.001	0.001	0.037
100 vs 150	0.122	0.189	0.631	0.497
100 vs 200	0.774	0.631	0.924	0.774
150 vs 200	0.189	0.071	0.924	0.275

Table 91: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0293, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0010.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.497
50 vs 150	0.774	0.001	0.001	0.378
50 vs 200	0.924	0.001	0.001	0.631
100 vs 150	0.631	0.631	0.016	0.774
100 vs 200	0.631	0.774	0.016	0.924
150 vs 200	0.631	0.497	0.189	0.497

Table 92: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0311, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.631	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.016
50 vs 150	0.924	0.001	0.001	0.774
50 vs 200	0.924	0.001	0.001	0.005
100 vs 150	0.924	0.378	0.189	0.016
100 vs 200	0.924	0.071	0.497	0.378
150 vs 200	0.924	0.378	0.071	0.001

Table 93: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0010, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.275	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.037	0.001	0.001	0.924
50 vs 150	0.122	0.001	0.001	0.924
50 vs 200	0.378	0.001	0.001	0.001
100 vs 150	0.001	0.924	0.924	0.774
100 vs 200	0.016	0.924	0.774	0.001
150 vs 200	0.016	0.924	0.774	0.001

Table 94: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0036, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0007.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.071	0.001	0.001	0.001
10 vs 100	0.071	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.275
50 vs 150	0.122	0.001	0.001	0.774
50 vs 200	0.631	0.001	0.001	0.378
100 vs 150	0.071	0.378	0.189	0.189
100 vs 200	0.275	0.001	0.001	0.774
150 vs 200	0.275	0.122	0.001	0.631

Table 95: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0027, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.005	0.001	0.001	0.001
10 vs 100	0.005	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.275	0.001	0.001	0.497
50 vs 150	0.189	0.001	0.001	0.016
50 vs 200	0.071	0.001	0.001	0.001
100 vs 150	0.924	0.275	0.189	0.001
100 vs 200	0.631	0.037	0.122	0.001
150 vs 200	0.924	0.001	0.001	0.497

Table 96: Pairwise Conover post-hoc p-values for RQ6 (Cloud Quantized Early Exit Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0029, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.016	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.122	0.001	0.001	0.001
50 vs 100	0.122	0.001	0.001	0.497
50 vs 150	0.631	0.001	0.001	0.122
50 vs 200	0.378	0.001	0.001	0.037
100 vs 150	0.071	0.631	0.774	0.016
100 vs 200	0.189	0.122	0.016	0.016
150 vs 200	0.497	0.001	0.001	0.631

Table 97: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0180, ResNext = 0.0183, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.774	0.631	0.001
10 vs 100	0.001	0.631	0.631	0.001
10 vs 150	0.001	0.275	0.924	0.001
10 vs 200	0.001	0.774	0.924	0.001
50 vs 100	0.001	0.631	0.924	0.001
50 vs 150	0.001	0.497	0.378	0.001
50 vs 200	0.001	0.497	0.189	0.001
100 vs 150	0.378	0.122	0.631	0.001
100 vs 200	0.631	0.924	0.275	0.001
150 vs 200	0.631	0.631	0.497	0.037

Table 98: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0198, ResNext = 0.0064, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.924	0.631	0.001
10 vs 100	0.001	0.774	0.071	0.001
10 vs 150	0.001	0.275	0.631	0.001
10 vs 200	0.001	0.189	0.378	0.001
50 vs 100	0.071	0.924	0.071	0.001
50 vs 150	0.001	0.774	0.378	0.001
50 vs 200	0.016	0.631	0.275	0.001
100 vs 150	0.189	0.774	0.122	0.001
100 vs 200	0.497	0.631	0.275	0.001
150 vs 200	0.122	0.275	0.631	0.001

Table 99: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0000, ResNet = 0.0023, ResNext = 0.0015, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.497	0.924	0.001
10 vs 100	0.001	0.122	0.001	0.001
10 vs 150	0.001	0.005	0.378	0.001
10 vs 200	0.001	0.122	0.774	0.001
50 vs 100	0.001	0.122	0.001	0.001
50 vs 150	0.001	0.016	0.378	0.001
50 vs 200	0.001	0.071	0.774	0.001
100 vs 150	0.122	0.774	0.122	0.001
100 vs 200	0.001	0.378	0.001	0.001
150 vs 200	0.001	0.631	0.631	0.005

Table 100: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0287, ResNext = 0.0096, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.924	0.924	0.001
10 vs 100	0.001	0.924	0.924	0.001
10 vs 150	0.001	0.631	0.497	0.001
10 vs 200	0.001	0.497	0.071	0.001
50 vs 100	0.001	0.774	0.774	0.001
50 vs 150	0.001	0.631	0.774	0.001
50 vs 200	0.001	0.924	0.189	0.001
100 vs 150	0.122	0.631	0.275	0.001
100 vs 200	0.071	0.924	0.037	0.001
150 vs 200	0.005	0.774	0.924	0.001

Table 101: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0036, ResNext = 0.0139, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.924	0.001
10 vs 100	0.001	0.497	0.189	0.001
10 vs 150	0.001	0.001	0.631	0.001
10 vs 200	0.001	0.071	0.924	0.001
50 vs 100	0.001	0.497	0.189	0.001
50 vs 150	0.001	0.774	0.378	0.001
50 vs 200	0.001	0.774	0.924	0.001
100 vs 150	0.122	0.378	0.378	0.001
100 vs 200	0.378	0.631	0.189	0.001
150 vs 200	0.275	0.774	0.924	0.005

Table 102: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0297, ResNext = 0.0293, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.924	0.378	0.001
10 vs 100	0.001	0.924	0.924	0.001
10 vs 150	0.001	0.924	0.924	0.001
10 vs 200	0.001	0.924	0.631	0.001
50 vs 100	0.001	0.774	0.774	0.001
50 vs 150	0.001	0.924	0.924	0.001
50 vs 200	0.001	0.497	0.631	0.001
100 vs 150	0.275	0.774	0.378	0.001
100 vs 200	0.275	0.774	0.774	0.001
150 vs 200	0.924	0.631	0.774	0.001

Table 103: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0006, ResNext = 0.0001, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.122	0.001	0.001
50 vs 150	0.001	0.378	0.378	0.001
50 vs 200	0.001	0.631	0.001	0.001
100 vs 150	0.189	0.497	0.037	0.001
100 vs 200	0.122	0.275	0.924	0.001
150 vs 200	0.071	0.497	0.016	0.001

Table 104: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0000, ResNet = 0.0002, ResNext = 0.0001, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.005	0.001	0.001
50 vs 150	0.001	0.001	0.001	0.001
50 vs 200	0.001	0.631	0.631	0.001
100 vs 150	0.005	0.122	0.924	0.001
100 vs 200	0.001	0.037	0.016	0.001
150 vs 200	0.001	0.378	0.037	0.001

Table 105: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0000, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.001	0.001
50 vs 150	0.001	0.001	0.001	0.001
50 vs 200	0.001	0.122	0.122	0.001
100 vs 150	0.001	0.924	0.774	0.001
100 vs 200	0.016	0.071	0.001	0.001
150 vs 200	0.001	0.001	0.005	0.001

Table 106: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0006, ResNext = 0.0001, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.122	0.497	0.001	0.001
50 vs 150	0.037	0.122	0.001	0.001
50 vs 200	0.497	0.275	0.122	0.001
100 vs 150	0.378	0.631	0.631	0.001
100 vs 200	0.001	0.631	0.037	0.001
150 vs 200	0.001	0.122	0.005	0.037

Table 107: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0002, ResNext = 0.0002, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.016	0.275	0.275	0.001
50 vs 150	0.001	0.001	0.001	0.001
50 vs 200	0.071	0.497	0.924	0.001
100 vs 150	0.001	0.631	0.275	0.001
100 vs 200	0.275	0.497	0.189	0.001
150 vs 200	0.001	0.001	0.001	0.037

Table 108: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Partition Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0001, ResNext = 0.0002, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.005	0.001
50 vs 150	0.001	0.001	0.016	0.001
50 vs 200	0.001	0.005	0.378	0.001
100 vs 150	0.497	0.924	0.378	0.001
100 vs 200	0.016	0.005	0.275	0.001
150 vs 200	0.005	0.016	0.037	0.037

Table 109: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0097, ResNext = 0.0023, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.710	0.323	0.000	0.040
1 vs 50	0.001	0.035	0.002	0.000
1 vs 100	0.000	0.002	1.000	0.000
1 vs 150	0.000	0.008	0.002	0.000
1 vs 200	0.000	0.049	0.002	0.000
10 vs 50	0.198	1.000	1.000	0.040
10 vs 100	0.002	0.595	0.033	0.000
10 vs 150	0.000	1.000	1.000	0.000
10 vs 200	0.000	1.000	1.000	0.000
50 vs 100	0.996	1.000	0.107	0.029
50 vs 150	0.006	1.000	1.000	0.000
50 vs 200	0.005	1.000	1.000	0.000
100 vs 150	0.597	1.000	0.156	0.000
100 vs 200	0.500	1.000	0.138	0.003
150 vs 200	1.000	1.000	1.000	1.000

Table 110: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0253, ResNext = 0.0011, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.594	0.038	0.041	0.001
1 vs 50	0.003	0.053	0.001	0.000
1 vs 100	0.000	0.149	0.000	0.000
1 vs 150	0.000	0.059	0.490	0.000
1 vs 200	0.000	0.013	0.000	0.000
10 vs 50	0.494	1.000	1.000	0.001
10 vs 100	0.000	1.000	0.714	0.000
10 vs 150	0.000	1.000	1.000	0.000
10 vs 200	0.000	1.000	0.063	0.000
50 vs 100	0.006	1.000	1.000	0.001
50 vs 150	0.012	1.000	0.255	0.000
50 vs 200	0.001	1.000	1.000	0.000
100 vs 150	1.000	1.000	0.063	0.001
100 vs 200	1.000	1.000	1.000	0.000
150 vs 200	1.000	1.000	0.004	0.001

Table 111: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0014, ResNext = 0.0024, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.050	0.392	0.159	0.001
1 vs 50	0.000	0.003	0.295	0.000
1 vs 100	0.000	0.000	0.000	0.000
1 vs 150	0.000	0.000	0.002	0.000
1 vs 200	0.000	0.046	0.000	0.000
10 vs 50	0.050	0.915	1.000	0.001
10 vs 100	0.000	0.009	0.423	0.000
10 vs 150	0.000	0.092	1.000	0.000
10 vs 200	0.000	1.000	0.205	0.000
50 vs 100	0.000	0.915	0.232	0.001
50 vs 150	0.000	1.000	0.749	0.000
50 vs 200	0.010	1.000	0.109	0.000
100 vs 150	0.128	1.000	1.000	0.001
100 vs 200	0.313	0.092	1.000	0.000
150 vs 200	0.000	0.724	1.000	0.001

Table 112: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0195, ResNext = 0.0131, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.390	0.023	0.075	0.001
1 vs 50	0.001	0.009	0.024	0.000
1 vs 100	0.000	0.050	0.002	0.000
1 vs 150	0.000	0.032	0.038	0.000
1 vs 200	0.000	0.267	0.179	0.000
10 vs 50	0.390	1.000	1.000	0.001
10 vs 100	0.000	1.000	1.000	0.000
10 vs 150	0.000	1.000	1.000	0.000
10 vs 200	0.000	1.000	1.000	0.000
50 vs 100	0.006	1.000	1.000	0.001
50 vs 150	0.004	1.000	1.000	0.000
50 vs 200	0.000	1.000	1.000	0.000
100 vs 150	1.000	1.000	1.000	0.001
100 vs 200	1.000	1.000	1.000	0.000
150 vs 200	1.000	1.000	1.000	0.001

Table 113: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0173, ResNext = 0.0081, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.208	0.021	0.013	0.001
1 vs 50	0.000	0.005	0.006	0.000
1 vs 100	0.000	0.122	0.747	0.000
1 vs 150	0.000	0.027	0.013	0.000
1 vs 200	0.000	0.168	0.004	0.000
10 vs 50	0.208	1.000	1.000	0.001
10 vs 100	0.000	1.000	1.000	0.000
10 vs 150	0.000	1.000	1.000	0.000
10 vs 200	0.000	1.000	1.000	0.000
50 vs 100	0.036	1.000	0.824	0.001
50 vs 150	0.000	1.000	1.000	0.000
50 vs 200	0.000	1.000	1.000	0.000
100 vs 150	0.127	1.000	1.000	0.001
100 vs 200	0.127	1.000	0.553	0.000
150 vs 200	1.000	1.000	1.000	0.001

Table 114: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0184, ResNext = 0.0234, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.870	0.067	0.017	0.001
1 vs 50	0.001	0.025	0.159	0.000
1 vs 100	0.000	0.141	0.024	0.000
1 vs 150	0.000	0.005	0.105	0.000
1 vs 200	0.000	0.074	0.026	0.000
10 vs 50	0.106	1.000	1.000	0.001
10 vs 100	0.008	1.000	1.000	0.000
10 vs 150	0.000	1.000	1.000	0.000
10 vs 200	0.000	1.000	1.000	0.000
50 vs 100	1.000	1.000	1.000	0.001
50 vs 150	0.033	1.000	1.000	0.000
50 vs 200	0.018	1.000	1.000	0.000
100 vs 150	0.376	1.000	1.000	0.001
100 vs 200	0.221	1.000	1.000	0.000
150 vs 200	1.000	1.000	1.000	0.001

Table 115: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0172, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0005.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.010	0.554	0.564	1.000
1 vs 50	0.099	0.003	0.003	0.000
1 vs 100	0.027	0.000	0.000	0.000
1 vs 150	0.228	0.000	0.000	0.000
1 vs 200	0.011	0.000	0.000	0.001
10 vs 50	1.000	0.554	0.564	0.000
10 vs 100	1.000	0.000	0.000	0.011
10 vs 150	1.000	0.000	0.000	0.011
10 vs 200	1.000	0.000	0.000	0.041
50 vs 100	1.000	0.003	0.003	0.779
50 vs 150	1.000	0.001	0.004	0.779
50 vs 200	1.000	0.004	0.002	0.241
100 vs 150	1.000	1.000	1.000	1.000
100 vs 200	1.000	1.000	1.000	1.000
150 vs 200	1.000	1.000	1.000	1.000

Table 116: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0045, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.380	0.378	0.387	0.963
1 vs 50	0.009	0.001	0.001	0.000
1 vs 100	0.094	0.000	0.000	0.000
1 vs 150	0.001	0.000	0.000	0.000
1 vs 200	0.001	0.000	0.000	0.000
10 vs 50	1.000	0.378	0.387	0.041
10 vs 100	1.000	0.000	0.000	0.000
10 vs 150	0.304	0.000	0.000	0.019
10 vs 200	0.340	0.000	0.000	0.000
50 vs 100	1.000	0.001	0.001	0.703
50 vs 150	1.000	0.000	0.000	1.000
50 vs 200	1.000	0.020	0.017	0.019
100 vs 150	1.000	1.000	1.000	1.000
100 vs 200	1.000	1.000	1.000	1.000
150 vs 200	1.000	0.845	0.861	0.041

Table 117: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0138, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0004.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.010	0.197	0.023	1.000
1 vs 50	0.070	0.000	0.000	0.003
1 vs 100	0.018	0.000	0.000	0.000
1 vs 150	0.006	0.000	0.000	0.000
1 vs 200	0.342	0.000	0.000	0.000
10 vs 50	1.000	0.197	0.023	0.205
10 vs 100	1.000	0.000	0.000	0.000
10 vs 150	1.000	0.000	0.000	0.000
10 vs 200	1.000	0.000	0.000	0.005
50 vs 100	1.000	0.000	0.000	0.148
50 vs 150	1.000	0.000	0.000	0.281
50 vs 200	1.000	0.033	0.023	1.000
100 vs 150	1.000	1.000	1.000	1.000
100 vs 200	1.000	0.252	0.001	1.000
150 vs 200	1.000	0.056	0.000	1.000

Table 118: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0028, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0004.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.363	0.251	1.000
1 vs 50	0.002	0.001	0.000	0.000
1 vs 100	0.001	0.000	0.000	0.000
1 vs 150	0.003	0.000	0.000	0.000
1 vs 200	1.000	0.000	0.000	0.001
10 vs 50	1.000	0.363	0.251	0.005
10 vs 100	1.000	0.000	0.000	0.007
10 vs 150	1.000	0.000	0.000	0.000
10 vs 200	0.083	0.000	0.000	0.085
50 vs 100	1.000	0.000	0.000	1.000
50 vs 150	1.000	0.000	0.001	1.000
50 vs 200	0.138	0.024	0.010	1.000
100 vs 150	1.000	1.000	0.760	1.000
100 vs 200	0.095	0.993	0.123	1.000
150 vs 200	0.226	1.000	1.000	0.118

Table 119: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0014, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0008.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.009	0.354	0.272	1.000
1 vs 50	0.000	0.001	0.001	0.002
1 vs 100	0.008	0.000	0.000	0.000
1 vs 150	0.000	0.000	0.000	0.000
1 vs 200	0.724	0.000	0.000	0.000
10 vs 50	1.000	0.354	0.272	0.100
10 vs 100	1.000	0.000	0.000	0.006
10 vs 150	0.915	0.000	0.000	0.003
10 vs 200	1.000	0.000	0.000	0.006
50 vs 100	1.000	0.001	0.000	1.000
50 vs 150	1.000	0.000	0.000	1.000
50 vs 200	0.017	0.018	0.031	1.000
100 vs 150	1.000	1.000	1.000	1.000
100 vs 200	1.000	1.000	0.341	1.000
150 vs 200	0.013	0.538	0.341	1.000

Table 120: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Partition Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0011, ResNet = 0.0002, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.002	0.840	0.120	1.000
1 vs 50	0.485	0.002	0.000	0.000
1 vs 100	0.000	0.000	0.000	0.001
1 vs 150	0.000	0.000	0.000	0.000
1 vs 200	0.040	0.000	0.000	0.000
10 vs 50	0.427	0.301	0.120	0.023
10 vs 100	1.000	0.000	0.000	0.069
10 vs 150	1.000	0.000	0.000	0.003
10 vs 200	1.000	0.001	0.000	0.000
50 vs 100	0.012	0.038	0.000	1.000
50 vs 150	0.012	0.004	0.000	1.000
50 vs 200	1.000	0.301	0.012	1.000
100 vs 150	1.000	1.000	0.344	1.000
100 vs 200	0.167	1.000	1.000	0.721
150 vs 200	0.167	1.000	0.005	1.000

Table 121: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0097, ResNext = 0.0023, DUC = 0.0001.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.774	0.631	0.001
10 vs 100	0.001	0.037	0.001	0.001
10 vs 150	0.001	0.122	0.774	0.001
10 vs 200	0.001	0.378	0.631	0.001
50 vs 100	0.037	0.631	0.016	0.001
50 vs 150	0.016	0.774	0.924	0.001
50 vs 200	0.016	0.924	0.774	0.001
100 vs 150	0.071	0.378	0.016	0.005
100 vs 200	0.071	0.189	0.016	0.001
150 vs 200	0.924	0.378	0.774	0.497

Table 122: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0253, ResNext = 0.0011, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.924	0.189	0.001
10 vs 100	0.001	0.631	0.275	0.001
10 vs 150	0.001	0.924	0.275	0.001
10 vs 200	0.001	0.924	0.001	0.001
50 vs 100	0.001	0.631	0.631	0.001
50 vs 150	0.001	0.924	0.016	0.001
50 vs 200	0.005	0.497	0.071	0.001
100 vs 150	0.774	0.924	0.071	0.001
100 vs 200	0.497	0.275	0.924	0.001
150 vs 200	0.497	0.631	0.001	0.001

Table 123: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0014, ResNext = 0.0024, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.071	0.924	0.001
10 vs 100	0.001	0.001	0.071	0.001
10 vs 150	0.001	0.016	0.275	0.001
10 vs 200	0.001	0.631	0.016	0.001
50 vs 100	0.001	0.071	0.071	0.001
50 vs 150	0.001	0.378	0.122	0.001
50 vs 200	0.001	0.378	0.001	0.001
100 vs 150	0.071	0.631	0.774	0.001
100 vs 200	0.122	0.016	0.924	0.001
150 vs 200	0.001	0.275	0.631	0.001

Table 124: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0195, ResNext = 0.0131, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.924	0.631	0.001
10 vs 100	0.001	0.924	0.275	0.001
10 vs 150	0.001	0.924	0.631	0.001
10 vs 200	0.001	0.378	0.631	0.001
50 vs 100	0.001	0.631	0.189	0.001
50 vs 150	0.001	0.497	0.774	0.001
50 vs 200	0.001	0.189	0.378	0.001
100 vs 150	0.774	0.774	0.378	0.001
100 vs 200	0.275	0.631	0.122	0.001
150 vs 200	0.189	0.378	0.774	0.001

Table 125: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0173, ResNext = 0.0081, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.631	0.924	0.001
10 vs 100	0.001	0.378	0.071	0.001
10 vs 150	0.001	0.924	0.924	0.001
10 vs 200	0.001	0.631	0.378	0.001
50 vs 100	0.001	0.378	0.037	0.001
50 vs 150	0.001	0.378	0.774	0.001
50 vs 200	0.001	0.275	0.924	0.001
100 vs 150	0.016	0.497	0.189	0.001
100 vs 200	0.122	0.774	0.122	0.001
150 vs 200	0.774	0.378	0.774	0.001

Table 126: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0184, ResNext = 0.0234, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.774	0.497	0.001
10 vs 100	0.001	0.924	0.774	0.001
10 vs 150	0.001	0.378	0.497	0.001
10 vs 200	0.001	0.924	0.924	0.001
50 vs 100	0.122	0.378	0.378	0.001
50 vs 150	0.071	0.497	0.924	0.001
50 vs 200	0.037	0.774	0.378	0.001
100 vs 150	0.001	0.189	0.631	0.001
100 vs 200	0.122	0.924	0.924	0.001
150 vs 200	0.497	0.378	0.774	0.001

Table 127: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0172, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0005.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.275	0.001	0.001	0.001
10 vs 100	0.924	0.001	0.001	0.001
10 vs 150	0.275	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.122
50 vs 150	0.497	0.001	0.001	0.122
50 vs 200	0.497	0.001	0.001	0.037
100 vs 150	0.497	0.774	0.924	0.924
100 vs 200	0.497	0.924	0.924	0.774
150 vs 200	0.378	0.774	0.924	0.631

Table 128: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0045, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.037	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.071	0.001	0.001	0.001
50 vs 100	0.378	0.001	0.001	0.189
50 vs 150	0.275	0.001	0.001	0.924
50 vs 200	0.275	0.001	0.001	0.001
100 vs 150	0.189	0.774	0.378	0.497
100 vs 200	0.122	0.378	0.275	0.631
150 vs 200	0.774	0.122	0.275	0.001

Table 129: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0138, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0004.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.497	0.001	0.001	0.001
10 vs 100	0.924	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.189	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.071
50 vs 150	0.631	0.001	0.001	0.071
50 vs 200	0.774	0.001	0.001	0.122
100 vs 150	0.497	0.497	0.497	0.774
100 vs 200	0.189	0.037	0.001	0.275
150 vs 200	0.122	0.071	0.001	0.378

Table 130: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0028, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0004.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.774	0.001	0.001	0.001
10 vs 150	0.497	0.001	0.001	0.001
10 vs 200	0.005	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.774
50 vs 150	0.924	0.001	0.001	0.189
50 vs 200	0.016	0.001	0.001	0.497
100 vs 150	0.774	0.924	0.071	0.189
100 vs 200	0.016	0.189	0.122	0.378
150 vs 200	0.016	0.189	0.378	0.016

Table 131: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0014, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0008.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.924	0.001	0.001	0.001
10 vs 150	0.037	0.001	0.001	0.001
10 vs 200	0.037	0.001	0.001	0.001
50 vs 100	0.189	0.001	0.001	0.497
50 vs 150	0.924	0.001	0.001	0.275
50 vs 200	0.005	0.001	0.001	0.275
100 vs 150	0.275	0.497	0.924	0.774
100 vs 200	0.189	0.378	0.071	0.774
150 vs 200	0.005	0.122	0.122	0.924

Table 132: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0011, ResNet = 0.0002, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.122	0.001	0.001	0.001
10 vs 100	0.378	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.189	0.001	0.001	0.001
50 vs 100	0.001	0.016	0.001	0.631
50 vs 150	0.037	0.001	0.001	0.497
50 vs 200	0.122	0.071	0.001	0.189
100 vs 150	0.774	0.631	0.122	0.275
100 vs 200	0.005	0.631	0.275	0.189
150 vs 200	0.071	0.189	0.005	0.378

Table 133: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0137, ResNext = 0.0080, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.378	0.924	0.001
10 vs 100	0.001	0.774	0.378	0.001
10 vs 150	0.001	0.774	0.275	0.001
10 vs 200	0.001	0.631	0.774	0.001
50 vs 100	0.001	0.071	0.378	0.001
50 vs 150	0.001	0.924	0.037	0.001
50 vs 200	0.001	0.774	0.497	0.001
100 vs 150	0.016	0.275	0.774	0.001
100 vs 200	0.037	0.189	0.275	0.001
150 vs 200	0.774	0.774	0.037	0.016

Table 134: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0242, ResNext = 0.0188, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.378	0.497	0.001
10 vs 100	0.001	0.924	0.497	0.001
10 vs 150	0.001	0.378	0.774	0.001
10 vs 200	0.001	0.924	0.189	0.001
50 vs 100	0.001	0.631	0.631	0.001
50 vs 150	0.001	0.924	0.774	0.001
50 vs 200	0.001	0.378	0.631	0.001
100 vs 150	0.001	0.924	0.378	0.001
100 vs 200	0.774	0.774	0.631	0.001
150 vs 200	0.037	0.631	0.497	0.001

Table 135: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0000, ResNet = 0.0010, ResNext = 0.0003, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.497	0.631	0.001
10 vs 100	0.001	0.037	0.001	0.001
10 vs 150	0.001	0.005	0.001	0.001
10 vs 200	0.001	0.005	0.924	0.001
50 vs 100	0.001	0.189	0.001	0.001
50 vs 150	0.001	0.016	0.016	0.001
50 vs 200	0.001	0.037	0.924	0.001
100 vs 150	0.001	0.924	0.378	0.001
100 vs 200	0.122	0.189	0.001	0.001
150 vs 200	0.001	0.497	0.005	0.037

Table 136: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0052, ResNext = 0.0065, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.774	0.378	0.001
10 vs 100	0.001	0.122	0.037	0.001
10 vs 150	0.001	0.189	0.378	0.001
10 vs 200	0.001	0.378	0.378	0.001
50 vs 100	0.001	0.631	0.037	0.001
50 vs 150	0.001	0.275	0.774	0.001
50 vs 200	0.001	0.774	0.924	0.001
100 vs 150	0.189	0.001	0.122	0.001
100 vs 200	0.037	0.924	0.378	0.001
150 vs 200	0.774	0.122	0.631	0.001

Table 137: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0170, ResNext = 0.0125, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.631	0.189	0.001
10 vs 100	0.001	0.774	0.631	0.001
10 vs 150	0.001	0.497	0.631	0.001
10 vs 200	0.001	0.631	0.016	0.001
50 vs 100	0.001	0.774	0.774	0.001
50 vs 150	0.001	0.497	0.774	0.001
50 vs 200	0.001	0.631	0.631	0.001
100 vs 150	0.016	0.275	0.774	0.001
100 vs 200	0.001	0.774	0.497	0.001
150 vs 200	0.774	0.122	0.497	0.189

Table 138: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Mobile-Edge BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0214, ResNext = 0.0218, DUC = 0.0000.

Edge-Cloud BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.275	0.924	0.001
10 vs 100	0.001	0.774	0.378	0.001
10 vs 150	0.001	0.631	0.497	0.001
10 vs 200	0.001	0.497	0.631	0.001
50 vs 100	0.001	0.497	0.497	0.001
50 vs 150	0.001	0.631	0.497	0.001
50 vs 200	0.001	0.378	0.631	0.001
100 vs 150	0.016	0.924	0.631	0.001
100 vs 200	0.001	0.924	0.497	0.001
150 vs 200	0.924	0.924	0.774	0.005

Table 139: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 1Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0081, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0005.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.378	0.001	0.001	0.001
50 vs 100	0.497	0.001	0.001	0.774
50 vs 150	0.071	0.001	0.001	0.924
50 vs 200	0.275	0.001	0.001	0.189
100 vs 150	0.378	0.378	0.071	0.497
100 vs 200	0.774	0.122	0.378	0.071
150 vs 200	0.631	0.275	0.189	0.071

Table 140: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 10Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0002, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.924	0.001	0.001	0.631
50 vs 150	0.016	0.001	0.001	0.122
50 vs 200	0.631	0.122	0.001	0.378
100 vs 150	0.071	0.774	0.378	0.189
100 vs 200	0.275	0.071	0.497	0.497
150 vs 200	0.005	0.001	0.037	0.774

Table 141: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 50Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0008, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.016	0.001	0.001	0.001
10 vs 100	0.037	0.001	0.001	0.001
10 vs 150	0.497	0.001	0.001	0.001
10 vs 200	0.189	0.001	0.001	0.001
50 vs 100	0.001	0.001	0.001	0.037
50 vs 150	0.071	0.001	0.001	0.071
50 vs 200	0.275	0.001	0.001	0.016
100 vs 150	0.275	0.774	0.497	0.378
100 vs 200	0.005	0.005	0.016	0.189
150 vs 200	0.189	0.001	0.189	0.275

Table 142: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 100Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0058, ResNet = 0.0001, ResNext = 0.0001, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.631	0.001	0.001	0.001
10 vs 150	0.497	0.001	0.001	0.001
10 vs 200	0.016	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.497
50 vs 150	0.378	0.001	0.001	0.122
50 vs 200	0.071	0.001	0.001	0.275
100 vs 150	0.774	0.774	0.122	0.037
100 vs 200	0.071	0.001	0.001	0.631
150 vs 200	0.189	0.016	0.071	0.037

Table 143: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 150Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0011, ResNet = 0.0001, ResNext = 0.0000, DUC = 0.0002.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.001	0.001	0.001
10 vs 100	0.378	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.071	0.001	0.001	0.001
50 vs 150	0.122	0.001	0.001	0.924
50 vs 200	0.001	0.001	0.001	0.924
100 vs 150	0.497	0.774	0.774	0.016
100 vs 200	0.071	0.071	0.001	0.037
150 vs 200	0.071	0.189	0.001	0.924

Table 144: Pairwise Conover post-hoc p-values for RQ6 (Edge-Cloud Quantized Early Exit Partition Models) with Edge-Cloud BW = 200Mbps

Kruskal-Wallis p-values for each model: FCN = 0.0045, ResNet = 0.0000, ResNext = 0.0001, DUC = 0.0006.

Mobile-Edge BW Pair	FCN	ResNet	ResNext	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.378	0.001	0.001	0.001
10 vs 150	0.071	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.631	0.001	0.001	0.924
50 vs 150	0.005	0.001	0.001	0.631
50 vs 200	0.924	0.001	0.001	0.122
100 vs 150	0.275	0.631	0.122	0.774
100 vs 200	0.497	0.001	0.037	0.071
150 vs 200	0.016	0.001	0.071	0.275

Table 145: Pairwise Conover post-hoc p-values for RQ6 (Mobile Identity Models) comparing Mobile Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0248, ResNet = 0.0008, ResNeXt = 0.0004, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.005
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.497	0.774	0.189	0.071
50 vs 150	0.774	0.774	0.122	0.924
50 vs 200	0.774	0.189	0.037	0.497
100 vs 150	0.378	0.774	0.774	0.016
100 vs 200	0.631	0.497	0.275	0.001
150 vs 200	0.497	0.378	0.497	0.774

Table 146: Pairwise Conover post-hoc p-values for RQ6 (Mobile Quantized Models) comparing Mobile Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0018, ResNet = 0.0013, ResNeXt = 0.0006, DUC = 0.0011.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.631	0.005	0.001	0.001
10 vs 100	0.497	0.005	0.001	0.001
10 vs 150	0.275	0.001	0.001	0.005
10 vs 200	0.774	0.001	0.001	0.016
50 vs 100	0.189	0.924	0.122	0.924
50 vs 150	0.001	0.631	0.189	0.189
50 vs 200	0.774	0.497	0.275	0.497
100 vs 150	0.001	0.774	0.924	0.189
100 vs 200	0.275	0.774	0.378	0.774
150 vs 200	0.001	0.275	0.631	0.924

Table 147: Pairwise Conover post-hoc p-values for RQ6 (Mobile Early Exit Models) comparing Mobile Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0004, ResNet = 0.0001, ResNeXt = 0.0002, DUC = 0.0003.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.497	0.001	0.001	0.001
10 vs 100	0.071	0.001	0.001	0.001
10 vs 150	0.774	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.071	0.001	0.924	0.005
50 vs 150	0.189	0.001	0.001	0.497
50 vs 200	0.001	0.001	0.122	0.924
100 vs 150	0.189	0.631	0.005	0.631
100 vs 200	0.001	0.774	0.071	0.005
150 vs 200	0.001	0.631	0.378	0.378

Table 148: Pairwise Conover post-hoc p-values for RQ6 (Mobile Quantized Early Exit Models) comparing Mobile Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0022, ResNet = 0.0008, ResNeXt = 0.0001, DUC = 0.0009.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.189	0.071	0.001	0.001
10 vs 100	0.189	0.001	0.001	0.005
10 vs 150	0.275	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.497	0.189	0.924	0.924
50 vs 150	0.005	0.275	0.001	0.122
50 vs 200	0.924	0.122	0.001	0.924
100 vs 150	0.005	0.924	0.122	0.189
100 vs 200	0.497	0.378	0.016	0.631
150 vs 200	0.037	0.497	0.122	0.774

Table 149: Pairwise Conover post-hoc p-values for RQ6 (Edge Identity Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0132, ResNet = 0.0003, ResNeXt = 0.0002, DUC = 0.0005.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.924	0.001	0.001	0.001
10 vs 100	0.378	0.001	0.001	0.001
10 vs 150	0.631	0.001	0.001	0.001
10 vs 200	0.275	0.001	0.001	0.001
50 vs 100	0.631	0.924	0.001	0.071
50 vs 150	0.378	0.071	0.122	0.378
50 vs 200	0.378	0.001	0.001	0.924
100 vs 150	0.378	0.275	0.924	0.497
100 vs 200	0.122	0.189	0.631	0.122
150 vs 200	0.631	0.924	0.924	0.378

Table 150: Pairwise Conover post-hoc p-values for RQ6 (Edge Quantized Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0250, ResNet = 0.0003, ResNeXt = 0.0002, DUC = 0.0008.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.497	0.001	0.001	0.001
10 vs 100	0.497	0.001	0.001	0.001
10 vs 150	0.497	0.001	0.001	0.001
10 vs 200	0.924	0.001	0.001	0.001
50 vs 100	0.774	0.016	0.016	0.497
50 vs 150	0.631	0.071	0.016	0.497
50 vs 200	0.774	0.016	0.016	0.631
100 vs 150	0.497	0.774	0.378	0.275
100 vs 200	0.631	0.774	0.275	0.497
150 vs 200	0.924	0.631	0.631	0.774

Table 151: Pairwise Conover post-hoc p-values for RQ6 (Edge Early Exit Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0142, ResNet = 0.0002, ResNeXt = 0.0002, DUC = 0.0005.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.497	0.001	0.001	0.005
10 vs 100	0.275	0.001	0.001	0.001
10 vs 150	0.924	0.001	0.001	0.001
10 vs 200	0.774	0.001	0.001	0.001
50 vs 100	0.774	0.001	0.001	0.631
50 vs 150	0.189	0.005	0.001	0.378
50 vs 200	0.497	0.071	0.037	0.122
100 vs 150	0.189	0.189	0.924	0.497
100 vs 200	0.378	0.497	0.378	0.122
150 vs 200	0.631	0.631	0.924	0.037

Table 152: Pairwise Conover post-hoc p-values for RQ6 (Edge Quantized Early Exit Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0145, ResNet = 0.0010, ResNeXt = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.774	0.001	0.001	0.001
10 vs 100	0.924	0.001	0.001	0.001
10 vs 150	0.122	0.001	0.001	0.001
10 vs 200	0.378	0.001	0.001	0.001
50 vs 100	0.924	0.924	0.001	0.122
50 vs 150	0.189	0.378	0.001	0.378
50 vs 200	0.924	0.378	0.001	0.001
100 vs 150	0.497	0.774	0.071	0.275
100 vs 200	0.924	0.774	0.275	0.071
150 vs 200	0.189	0.924	0.924	0.001

Table 153: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Edge Partition Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0003, ResNeXt = 0.0001, DUC = 0.0000.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.071	0.016	0.001
50 vs 150	0.001	0.016	0.037	0.001
50 vs 200	0.001	0.071	0.001	0.001
100 vs 150	0.378	0.774	0.378	0.001
100 vs 200	0.275	0.631	0.001	0.001
150 vs 200	0.924	0.378	0.378	0.001

Table 154: Pairwise Conover post-hoc p-values for RQ6 (Mobile-Edge Quantized Early Exit Partition Models) comparing Mobile-Edge Bandwidths

Kruskal-Wallis p-values for each model: FCN = 0.0001, ResNet = 0.0002, ResNeXt = 0.0001, DUC = 0.0001.

Mobile-Edge BW Pair	FCN	ResNet	ResNeXt	DUC
1 vs 10	0.001	0.001	0.001	0.001
1 vs 50	0.001	0.001	0.001	0.001
1 vs 100	0.001	0.001	0.001	0.001
1 vs 150	0.001	0.001	0.001	0.001
1 vs 200	0.001	0.001	0.001	0.001
10 vs 50	0.001	0.001	0.001	0.001
10 vs 100	0.001	0.001	0.001	0.001
10 vs 150	0.001	0.001	0.001	0.001
10 vs 200	0.001	0.001	0.001	0.001
50 vs 100	0.001	0.037	0.001	0.001
50 vs 150	0.001	0.016	0.001	0.001
50 vs 200	0.001	0.001	0.001	0.001
100 vs 150	0.275	0.378	0.497	0.924
100 vs 200	0.037	0.378	0.924	0.005
150 vs 200	0.378	0.924	0.774	0.001