

# QPART: Adaptive Model Quantization and Dynamic Workload Balancing for Accuracy-aware Edge Inference

Xiangchen Li, *Student Member, IEEE*, Saeid Ghafouri, Bo Ji, *Senior Member, IEEE*, Hans Vandierendonck, *Senior Member, IEEE*, Deepu John, *Senior Member, IEEE*, Dimitrios S. Nikolopoulos, *Fellow, IEEE*,

**Abstract**—As machine learning inferences increasingly move to edge devices, adapting to diverse computational capabilities, hardware, and memory constraints becomes more critical. Instead of relying on a pre-trained model fixed for all future inference queries across diverse edge devices, we argue that planning an inference pattern with a request-specific model tailored to the device’s computational capacity, accuracy requirements, and time constraints is more cost-efficient and robust to diverse scenarios. To this end, we propose an accuracy-aware and workload-balanced inference system that integrates joint model quantization and inference partitioning. In this approach, the server dynamically responds to inference queries by sending a quantized model and adaptively sharing the inference workload with the device. Meanwhile, the device’s computational power, channel capacity, and accuracy requirements are considered when deciding.

Furthermore, we introduce a new optimization framework for the inference system, incorporating joint model quantization and partitioning. Our approach optimizes layer-wise quantization bit width and partition points to minimize time consumption and cost while accounting for varying accuracy requirements of tasks through an accuracy degradation metric in our optimization model. To our knowledge, this work represents the first exploration of optimizing quantization layer-wise bit-width in the inference serving system, by introducing theoretical measurement of accuracy degradation. Simulation results demonstrate a substantial reduction in overall time and power consumption, with computation payloads decreasing by over 80% and accuracy degradation kept below 1%.

**Index Terms**—Adaptive Inference Service, Model Quantization, Partitioning

## I. INTRODUCTION

ALTHOUGH deep learning has had tremendous success in computer vision [1] [2], natural language processing [3] [4] and autonomous transportation [5], the deployment of these already-trained and frontier models is still performed on server clusters [6], making the edge users exposed to the limitation of high latency [7], increased dependency on high transmission bandwidth [8], and potential privacy risks due to data transmission to centralized servers [7]. Many earlier efforts attempted to bring the frontier pre-trained machine learning models to the network’s edge. For instance, dedicated hardware accelerators like Google’s Edge TPU and Nvidia’s Jetson are optimized for performing neural network inference on edge devices [9] [10]. Within the NAS (Network Structure

Search) realm, several efforts have tried to discover neural networks optimized for deployment on edge devices [11] [12]. Model compression is a common and low-cost method to deploy neural networks. Many schemes and frameworks, such as early exit [13] [14], model pruning [15] [16], autoencoders [17], and knowledge distillation [18], have been rigorously explored and integrated into the inference offloading problem within the distributed network. From the perspective of computation, inference offloading and model partitioning [19] [20] have been promising tools for reducing the cost of deploying the machine learning models on the edge by offloading partial or all of the computation to more powerful servers.

While the performance and potential of certain technological advancements facilitating machine learning models on edge devices have been extensively explored in numerous prior studies, their constraints—such as hardware installation costs, model retraining expenses, and the costs associated with fine-tuning—remain significant considerations. Moreover, when deploying machine learning tasks on edge devices, the device type, memory footprint, computation capacity, expected accuracy, and channel conditions impact the inference performance and evolve over time, making it inappropriate to find a universal solution across all future inference queries. Specifically, there are a few factors to be considered before pushing the inference toward the edge of the network:

- 1) **Different Hardware Limitations on the Edge:** Edge devices like cell phones, smart watches, network cameras, and VR/AR glasses generate machine learning inference queries but differ in computational and memory capabilities. Different computation capacity comes from the variations in processor architecture, instruction sets, and clock rates [21] [22] [23]. Devices with powerful processors are more capable of handling inference locally. At the same time, those with low power or limited memory usually use methods like partitioning, quantization, pruning, or distillation to reduce workload. Moreover, the memory capacity also prevents large models to be fully deployed on the edge.
- 2) **Changing Transmission Conditions:** The wireless transmission channel significantly impacts inference service [24]. According to Shannon’s capacity formula [25], higher SNR leads to greater bandwidth and lower latency. Hence, higher accuracy may be achieved by

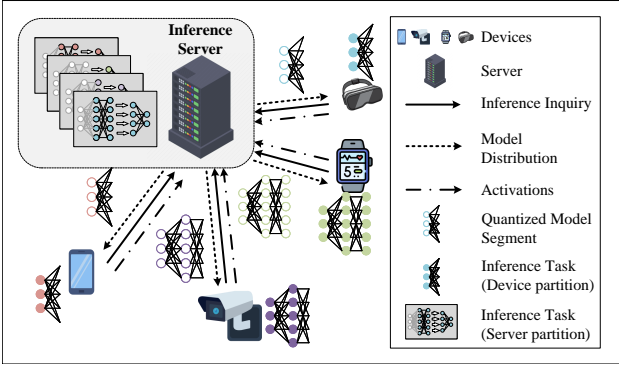


Fig. 1. Layer-wise Parameter Size Reduction This figure explains how the *QPART* serves the machine learning inference for edge devices. First, an inference request is initialized by edge devices and sent to the server. Second, the server responds to the edge devices by returning a quantized model segment. Then, edge devices process the local inference and feedback on the activation for server inference.

trading with latency. Considering channel capacity in inference or offloading makes solutions more robust against network fluctuations.

- 3) **Diverse Accuracy Requirement for Applications:** Desired accuracy varies by application. Critical tasks like medical diagnosis [26], autonomous vehicles [5], and power grid monitoring [27] demand high accuracy. In contrast, applications like recommendation systems [28] and social media [29] tolerate some accuracy loss. Accepting minor accuracy degradation can often save significant computing resources [30] [31].

To facilitate various devices on the edge with the already-trained neural network models without huge cost, this paper proposes an accuracy-aware and model-less inference system that responds to the inference request from edge devices by automatically quantizing the model for the devices and sharing the computation workload with the device. Shown in Fig.1, as a machine learning inference infrastructure, our proposed system *QPART* maintains the pre-trained machine learning models for a set of edge applications on the cloud server. Once the inference task has been generated from the edge devices, the inference request will be sent to the *QPART* specifying the type of the task, the channel capacity and the computation capacity of the device, and the accuracy requirement. Once the *QPART* receives the request from edge devices, it solves the joint optimization problem constrained by the accuracy degradation, which is modeled as a layer-wise accuracy degradation measurement, to find the model partition point and layer-wise quantization bit-width according to the current situation. After that, the first part of the whole neural network before the partition point, which is referred as model segment in this paper, will be quantized layer-wise (to reduce the transmission latency and the memory footprint on the device) and then be transmitted from the server to the device for the first stage inference. After the raw data collected by the edge device has been forwarded through the first part of the neural network model on the device, the intermediate output will be transmitted back to the *QPART* server for the consecutive inference. Finally, the inference output will be

sent back to the device. Our experimental evidence indicates that our framework can achieve up to an 80% reduction in communication payload, which contributes the most of the latency in a broad range of edge computing settings [32], while limiting accuracy degradation to below 1%. A comparative analysis with existing model compression methods further validates the efficiency of our approach.

The contributions of this work are summarized as follows:

- 1) *QPART*, an inference serving system consisting of adaptive model quantization and automatic computation workload balancing was proposed. Instead of scaling the pre-trained ML models for the edge manually, inference queries from the edge device are responded by *QPART* with optimized model segment for lower latency, energy consumption and cost, also with guaranteed accuracy.
- 2) A joint optimization framework of layer-wise quantization and model partitioning is built for the *QPART* system. The computation workload, time consumption, energy consumption, the cost of servers' resources, and transmission latency are modeled within the inference request, forming an optimization problem where the accuracy requirement and memory capacities are modeled as two constraints.
- 3) The proposed optimization problem is well-solved, and the close-form solution of quantization bit-width and partition point are derived. Further, the offline quantization algorithm and online scheduling algorithm are proposed to optimize the partition point and layer-wise quantization in runtime.
- 4) By introducing accuracy degradation measurement [33] based on quantization noise, the optimization problem is solved, enabling *QPART* to optimize quantization bit-width and partition point without violating the accuracy requirement.
- 5) A simulation platform is built, and numerous experimental results are carried out to show the efficiency of the proposed *QPART* serving inference queries.

The paper is organized as follows: Section II reviews the related work of inference offloading and model scaling. Section III outlines the target *QPART* system settings and the modeling of inference task offloading, including model quantization and partitioning. Section IV details the optimization problem's formulation and the corresponding algorithm's introduction. The execution and discussion of simulations are covered in Section V, and Section VI provides the conclusion.

## II. BACKGROUND AND RELATED WORK

As an inference serving system, *QPART* features model quantization and distributed inference, which makes the system located at the intersection of model scaling and inference offloading methods. This section will start by elaborating on these two fundamental techniques that allow the pre-trained model to be deployed on the edge.

### A. Deploying Models towards the Edge

Many modified neural networks or novel learning paradigms have been proposed to address the imbalance between mobile

TABLE I  
COMPARISON OF RELATED SYSTEMS ACCOMMODATING THE NEURAL NETWORKS ON THE EDGE; LATENCY: IS TIME CONSUMPTION CONSIDERED AND OPTIMIZED?; ENERGY: IS POWER CONSUMPTION CONSIDERED AND OPTIMIZED?; ACCURACY: IS THE INFERENCE ACCURACY CONSIDERED?; SCALING: DOES THE SYSTEM SCALE THE MODEL FOR THE EDGE DEVICES?; OFFLOADING: DOES THE SYSTEM SUPPORT INFERENCE OFFLOADING FROM THE DEVICES?

System	Latency	Energy	Accuracy	Scaling	Offloading
ApproRR [34]	✓	✓	✗	✗	✓
DeepCOD [35]	✓	✗	✓	✗	✓
DINA-P [36]	✓	✗	✗	✗	✓
COB [37]	✓	✗	✗	✗	✓
Ly-EXP4 [38]	✓	✗	✓	✗	✓
MASITO [39]	✓	✓	✓	✗	✓
INFaaS [40]	✗	✗	✓	✓	✗
FANN-on-MCU [41]	✗	✓	✓	✓	✗
PCA-driven Hybrid-Net [42]	✗	✓	✗	✓	✗
<i>QPART</i>	✓	✓	✓	✓	✓

devices and server computational resources. Early exit, model pruning, and auto-encoder, as techniques designed to enhance the computational efficiency of deep neural networks, are beneficial for devices with limited resources. Barbosa et al. [43] propose the Decision Early Exit (DEEX) approach, which is implemented at a network's first early exit point. DEEX aims to reduce total inference time by avoiding unnecessary evaluations at early exits that do not enhance model performance. Ren et al. [44] that the proposed offloading mechanism, aided by model pruning, can significantly reduce latency and energy consumption compared to other algorithms. By proposing a two-step pruning framework for partitioning DNNs between mobile devices and MEC servers, Shi et al.'s work [45] effectively reduced the wireless transmission workload of the device or the total computation workload, highlighting the benefits of model pruning in cooperative inference between devices and MEC servers. Further, auto-encoders were also utilized to facilitate inference on the edge. For instance, Hu et al. [17] investigated the streaming inference of CNN to simultaneously improve throughput and accuracy by communication compression. Specifically, an auto-encoder was inserted into the CNN model to compress the communication payload.

Many machine learning-based inference offloading algorithms effectively reduce latency and maintain accuracy, but they often require modifications to the base model, hindering scalability and adaptability. For instance, early exit mechanisms [46] [43] [47] necessitate additional training for each exit point, increasing training costs. Model pruning methods [44] [45] risk accuracy loss when applied across various neural networks, and efforts to mitigate these issues, such as re-tuning, are labor-intensive. Integrating autoencoders [17] adds unnecessary computational overhead. In contrast, our proposed model mitigates the computational burden by quantizing model parameters and intermediate outputs to reduced bit widths, thus circumventing additional computations or modifications to the original neural networks.

### B. Offloading the Inference Workload on Servers

Inference offloading, the process of transferring data-intensive and computation-heavy tasks from resource-

constrained devices to more capable computing nodes, faces significant bandwidth, latency, and energy consumption challenges.

Latency is a critical focus in inference offloading. Kang et al. [19] introduced Neurosurgeon, which reduces inference latency, power consumption, and system throughput by partitioning computation across edge devices and cloud servers. Duan et al. [48] extended this by jointly optimizing DNN partitioning and pipeline scheduling to lower latency further. Zhao et al. [49] explored the trade-off between latency and accuracy in edge-cloud offloading, focusing on offloading raw data with reduced quality. The challenge of power efficiency in offloading due to limited power on edge devices has also been studied. Xu et al. [34] optimized device and server energy consumption in 5G-enabled mobile edge clouds. Sada et al. [50] set latency and energy consumption as constraints while optimizing inference accuracy, but their approach required maintaining multiple model variants on the edge, which is impractical.

In this paper, we proposed *QPART*, which, to our best knowledge, is the first inference serving platform that combines automatic model quantization and adaptive inference offloading. By quantizing the model and optimizing the computation offloading pattern on the server, no model variants need to be pre-tuned or stored.

Proposed by Remero et al. [40], INFaaS is an automated model-less inference serving system, which is the closest related work to *QPART*. In their work, the model variant searching system [40] responds to the inference queries from a specific platform with an accuracy requirement by searching the optimal neural network variant in the set of scaled neural network models. The developer can specify their application's performance and accuracy requirements without specifying any model-specific parameters or structures. However, INFaaS is only a model variant searching system that is unsuitable for edge devices, and it only tries to search for and respond with the optimal model variant picked from a particular set. Our proposed *QPART* system is an end-to-end inference serving platform for edge computing that scales the model and shares the inference workload adaptively. Moreover, to reduce the transmission delay and save the memory footprint on the device, we solve the layer-wise quantization bit-width with close-form solutions instead of manually searching for the optimal model variant within a limited number of fixed model variants pre-defined. Also, NFaaS is not designed for practical edge computing scenarios: It does not consider the device's computation capacity or channel capacity. Table I provides a comparison summary between *QPART* and related works.

## III. SYSTEM MODEL

In this section, we elaborate on our proposed inference serving system *QPART* from the perspective of system overview, inference offloading design, model quantization design, and communication model.

### A. MEC system

First, we start from the whole picture of our proposed system shown in Fig. 2. Within our proposed inference system,

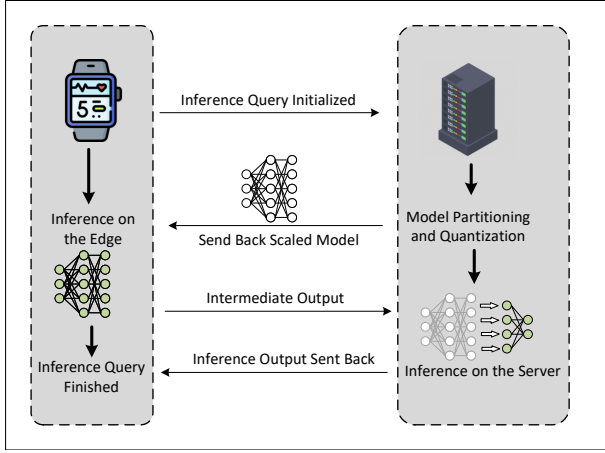


Fig. 2. Data Flow between Edge Devices and QPART System

multiple distinct inference requests will be initiated from edge devices, each characterized by varying neural network structures. We denote an inference query  $r$  from a mobile device in the time slot  $t$  by a tuple  $r = (\theta, a)$ , where  $\theta$  indicates the specific neural network model and  $a$  is the accuracy demand, which is defined as the maximal acceptable accuracy degradation compared to the initial accuracy of the model. Once the server receives the inference query, the optimized layer-wise quantization bit-width and model partition point will be solved by the QPART system. Specifically, model quantization, without degrading the prediction accuracy beyond requirements, effectively decreases the communication payload, thereby reducing communication latency, a primary contributor to overall inference latency in MEC networks [32]. We split an inference request into two sequential segments through model partitioning, which is processed on the edge device and the server, respectively. Further, the first model segment will be quantized and then sent back to the edge device. The edge device will perform the inference with a quantized model segment and send the intermediate output to the server for the inference of second model segment. Once all inference has been operated, the final results will be returned to the device or other nodes if needed.

### B. Inference Offloading

As illustrated in Fig. 2, each inference query  $r$  originating from an edge device will be divided into two distinct segments for distributed inference on the device and the server, respectively. We will elaborate on the model partition process and analyze the elements contributing to time and energy consumption, including device-based and server-based inferences and communication-related overhead.

For request  $r$  with network model  $\theta$ , the activation size of layer  $l \in \{1, 2, \dots, L\}$  is indicated as  $z_l^x$ , the number of weight parameters in layer  $l$  is  $z_l^w$ . Additionally, we quantify the computational demand of layer  $l$  in terms of multiply-accumulate (MAC) operations, referred to as  $o(l)$ . To adapt our model to various scenarios, we will present the formulation of  $o(l)$  for both linear (fully connected) and convolutional layers.

1) *Linear Layer Computation Cost*: If layer  $l$  is a linear layer with input activation  $X \in \mathbb{R}^{1 \times D}$  and weight matrix  $W \in \mathbb{R}^{D \times G}$ , the output activation can be expressed as  $Y = X \cdot W$ ,  $Y \in \mathbb{R}^{1 \times G}$ . Hence, the computation cost of layer  $l$  can be expressed in Eq. 1 [51]:

$$o(l) = D \times G \quad (1)$$

2) *Convolutional Layer Computation Cost*: A standard convolutional layer has the computation cost of [51]:

$$o(l) = C_{in} \times C_{out} \times F_1 \times F_2 \times U \times V \quad (2)$$

where  $C_{in}$  and  $C_{out}$  are the numbers of input channels, and output channels of layer  $l$ ,  $F_1$  and  $F_2$  define the size of the convolutional filter, and  $U$  and  $V$  are the size of input activation.

Therefore, with the optimal model partition point given  $p$  for request  $r$ , the computational costs of two segments of the neural network  $s$  can be formulated in Eq. 3 and Eq. 4:

$$O_1(p) = \sum_{l=1}^{p-1} o(l) \quad (3)$$

$$O_2(p) = \sum_{l=p}^L o(l) \quad (4)$$

Let  $f^{local} > 0$  denote the clock rate of the edge device when inferring the first part of the request. The time consumption of local inference can be expressed as:

$$T_{local} = O_1 \cdot \gamma_{local} \cdot \frac{1}{f^{local}} \quad (5)$$

where  $\gamma_{local}$  is a constant, referring to the average number of clock cycles needed to finish one MAC operation on the mobile user device. Generally, it depends on the processor architecture, instruction set architecture, pipeline, etc.

According to [52], the power consumption of a processor chip is proportional to  $V^2 \cdot f_{clock}$ , where  $V$  represents the circuit supply voltage, and  $f_{clock}$  is the clock rate. Furthermore, [53] notes that the clock frequency is approximately linearly proportional to the voltage supply, particularly when operating at low voltage limits. Therefore, the energy consumption per clock cycle can be defined as  $\kappa f_{clock}^2$ , where  $\kappa$  denotes the energy efficiency parameter, which depends on the hardware architecture. Consequently, the energy consumption for local inference can be formulated as follows.

$$E_{local} = \kappa (f^{local})^2 \cdot O_1 \cdot \gamma_{local} \quad (6)$$

where  $\kappa_n$  is the energy efficiency parameter of device  $n$ . In a similar vein, the subsequent expression of time consumption of server-based inference can be obtained:

$$T_{server} = O_2 \cdot \gamma_{server} \cdot \frac{1}{f_{server}} \quad (7)$$

Inference's energy consumption on the server side is not part of our model due to the data center's continuous power supply.

As a inference serving platform, QPART not only tailors the model variant for the edge devices, but also shares the



computation workload with the devices, by forwarding the second model segment on the server. The edge devices are supposed to be charged for the computation resources on the server. The cost for occupying the computation resources on the server per second is denoted by  $\zeta$ . Hence, the overall cost for inferencing the second segment on the server can be defined as Eq.8:

$$C = O_2 \cdot \gamma_{server} \cdot \frac{\zeta}{f_{server}} \quad (8)$$

### C. Model Quantization

Given the prevalence of hardware limitations in resource-constrained scenarios, such as embedded systems and wearable devices [54], coupled with the fact that many current neural network models are heavily overparameterized [55], the focus on training and inference with limited precision has garnered increasing attention among researchers [56] [57]. Model quantization, which involves reducing the precision of weights and biases in machine learning models, especially deep neural networks (DNNs), has emerged as a key technique. This process is integrated into our proposed *QPART* because it significantly reduces a neural network's memory footprint, facilitating its deployment on resource-constrained devices and reducing communication latency.

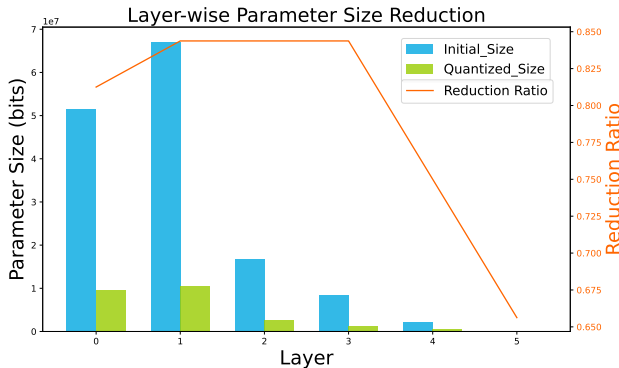


Fig. 3. Layer-wise Parameter Size Reduction. From the figure, a substantial part of the model memory footprint can be saved by adopting the quantization method in *QPART*.

Fig. 3 shows the layerwise parameter reduction achieved by implementing our proposed algorithm. Across all layers, the sizes of the parameters are substantially reduced by 62% to 84%, averaging 77%, which is referred to the reduction ratio in the figure, while the accuracy degradation remains below 1%. This figure proves the effectiveness of model quantization in inference offloading.

We will specifically focus on post-training quantization, which involves quantizing a model after it has been fully trained with full precision. By adopting this approach, our proposed quantized edge inference scheme can be seamlessly integrated into most existing edge inference applications without necessitating the retraining of deep learning models. In the subsequent parts of this section, we will introduce the notations used for quantization and quantized models and activations.

For an request  $r$ , we assume that the corresponding neural network model  $\theta$  consists of  $L$  learnable layers in total and that every layer can be denoted as layer  $l, l \in \{1, 2, \dots, L\}$ . The trained model parameters  $\theta$  are stored with double floating point precision. To lower the precision of the model's parameters and the intermediate activation to a more concise format while ensuring that the model's accuracy or generalization capability remains largely unaffected, we first define the uniform asymmetric quantizer as follows:

Given a real value  $c$ , which can be a weight or activation in neural networks, the quantized value using  $b$  bits  $c^q$  is selected from the following quantization set  $\mathcal{Q}$ :

$$\mathcal{Q} = [\mu : \frac{1}{2^b - 1} : \phi] + q^z \quad (9)$$

where  $\mu$  and  $\phi$  define the range of quantization set  $\mathcal{Q}$ ,  $[\mu : \frac{1}{2^b - 1} : \phi]$  denotes the uniform grid on  $[\mu, \phi]$  with interval  $\frac{1}{2^b - 1}$  between adjacent elements,  $q^z$  is the zero point. We define  $Q(\cdot)$  as the quantization function, therefore the quantized value  $c^q$  is given by Eq.10:

$$c^q = Q(c, b, q_z) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} (c - q) \quad (10)$$

This paper adopts layer-wise quantization and model partitioning techniques to serve the inference request for edge devices. We employ the notation  $p \in 1, 2, \dots, L$  to signify the partition point within the model, which serves as a strategic division for processing under our proposed optimization framework. The first neural network segment will be returned to the edge device for local inference. Furthermore, the quantization bit width for each layer in the first neural network segment is defined by a quantization bit-width vector,  $\mathbf{b} = [b_i]$ , with  $i \in 1, 2, \dots, p$ . This vector  $\mathbf{b}$  is essential as it determines the quantization granularity for each layer, thereby influencing the model's computational efficiency and overall performance.

### D. Communication Model

Our proposed MEC system uses a standard wireless communication framework to assess transmission latency. Once the model has been quantized and partitioned for the inference request on the server, the quantized model segment will be transmitted to the edge device through a wireless channel. Also, the intermediate output of the model segment will be transmitted from the device to the server after local inference is completed. Assuming that the transmission time is comparable to the channel coherence period [58], the channel gain of the wireless link from the mobile device to server can be defined as Eq.11:

$$g = \alpha \cdot h \quad (11)$$

where  $h$  is the small-scale fading component, which is frequency-dependent and exponentially distributed with unit mean, and  $\alpha$  is the large-scale fading component, which is frequency-independent, reflecting path loss and shadowing. Let  $\pi$  denote the transmit power of device, we will have the received SNR shown in Eq. 12:

$$\beta_k = \frac{\pi \cdot g}{\sigma} \quad (12)$$

Therefore, with a given channel bandwidth  $B$ , the channel capacity [25] can be formulated in the equation. 13:

$$r = B \cdot \log_2(1 + \frac{\pi \cdot g}{\sigma}) \quad (13)$$

As for the quantized model parameters and activations to be transmitted, which make up most of the communication payload, we can easily have their data size (in bits) as follows:

$$Z = Z_x + Z_w = b_p \cdot z_p^x + \sum_{l=1}^p b_l \cdot z_l^w \quad (14)$$

where  $Z_x$  and  $Z_w$  are the activation and model parameters communication payload, respectively. Finally, we have the transmission latency  $T_{tran}$  and energy consumption  $E_{tran}$  of the inference request  $r_n^t$  expressed in Eq. 15 and Eq. 16:

$$T_{tran} = \frac{Z}{r} \quad (15)$$

$$E_{tran} = \pi \cdot \frac{Z}{r} \quad (16)$$

#### IV. PROBLEM FORMULATION AND PROPOSED ALGORITHM

In the inference offloading scenario we consider, each user generates inference requests. Our primary objective is to develop an optimal offloading policy that minimizes inference latency and devices' power consumption while ensuring the accuracy of each inference. Specifically, the optimization problem we aim to solve can be formulated as follows:

$$\begin{aligned} \min_{b_i, p_n^t} \quad & \omega \cdot (T_{local} + T_{tran} + T_{server}) + \\ & \tau \cdot (E_{local} + E_{tran}) + \eta \cdot C \\ \text{s.t.} \quad & acc_{\theta} - acc_{\theta'} \leq a \end{aligned} \quad (17)$$

where  $\omega, \tau, \eta \geq 0$  are the significance weights for time consumption, energy consumption and cost, respectively. The weights offer the flexibility for each mobile device to tailor its preferences, striking a balance between performances and costs. Further, by letting some weights being zero, the objective can be altered towards different expectations. For example, should a mobile device  $n$  exhibit a heightened sensitivity to inference latency coupled with a long battery life, increasing the value of  $\omega$  would be beneficial.

Inspired by Yiren [33], we first define quantization noise and adversarial noise to quantify the impact of accuracy requirement  $a_n^t$  on the objective function. Specifically, we define  $\sigma_l^w$  and  $\sigma_p^x$  as the quantization noises on the last activation that are caused by the quantization of  $l$ -th layer's weights and activation, respectively, and they can be modeled as follows [33]:

$$\|\sigma_l^w\|_2^2 = s_l \cdot e^{-ln4b_l} \quad (18)$$

$$\|\sigma_p^x\|_2^2 = s_p \cdot e^{-ln4b_p} \quad (19)$$

Based on the noises we defined above, two measurements estimating the accuracy degradation caused by quantization of weights and activation in layer  $l$  can be defined [33], shown in Eq. 20 and Eq. 21:

$$\psi_l^w = \frac{\|\sigma_l^w\|_2^2}{\rho_l(a)} \quad (20)$$

$$\psi_l^x = \frac{\|\sigma_l^x\|_2^2}{\rho_l(a)} \quad (21)$$

where  $\rho_l(a_n^t)$  is the robustness parameter of layer  $l$ , and can be calculated with given dataset  $\mathcal{D}$  by Eq. 22 [33]:

$$\rho_l(a) = \frac{\text{mean}(\sigma_l^w, \sigma_l^x)}{\text{mean}(\sigma^*)} = \frac{\frac{1}{2|\mathcal{D}|} \cdot \sum_{x \in \mathcal{D}} (\|\sigma_l^w\|_2^2 + \|\sigma_l^x\|_2^2)}{\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \|\sigma^*\|_2^2} \quad (22)$$

where  $\sigma^*$  is the adversarial noise on the last layer's activation, which is the minimum noise that can change the classifying result of the neural network. As the measurement of the degradation of the accuracy caused by the quantification in layers,  $\psi_l^w = \psi_{l'}^w$  indicates that quantification at layers  $l$  and  $l'$  affects the accuracy of the model equally. Given that the linearity and additivity of this metric have already been established, and considering the independence of these measurements across all layers, we can reformulate the optimization problem originally presented in Equation 17 into a new form, as shown in Eq 23:

$$\begin{aligned} \min_{b, p} \quad & \xi \cdot O_1 + \delta \cdot O_2 + \epsilon \cdot (b_p \cdot z_p^x + \sum_{l=p}^L b_l \cdot z_l^w) \\ \text{s.t.} \quad & \psi_p^x + \sum_{l=p}^N \psi_l^w \leq \Delta \end{aligned} \quad (23)$$

where  $\Delta$  is a constant related to model accuracy degradation  $a$ , specifically, a higher  $\Delta$  means a higher  $a$ . Additionally, local cost  $\xi$ , server cost  $\delta$ , and transmission cost  $\epsilon$  are defined for simplification, and they can be detailed as follows:

$$\xi = \frac{\omega \cdot \gamma_{local}}{f_{local}} + \tau \gamma_{local} \kappa_n (f_{local}^n)^2 \quad (24)$$

$$\delta = \frac{(\omega + \eta \xi) \cdot \gamma_{server}}{f_{server}} + \tau \gamma_{server} \eta_m (f_{server}^n)^2 \quad (25)$$

$$\epsilon = \frac{\omega + \pi_n \tau}{r_m^n} \quad (26)$$

Then, the optimal value of quantization bit-width  $b$  and partition point  $p$  can be derived when:

$$\begin{aligned} \frac{z_p \rho_p}{s_p e^{-ln4b_p}} &= \frac{z_{p+1} \rho_{p+1}}{s_{p+1} e^{-ln4b_{p+1}}} = \dots = \frac{z_{N+1} \rho_{N+1}}{s_{N+1} e^{-ln4b_{N+1}}} \\ &= \frac{ln4 \cdot [\xi o(p) - \delta o(p) - \epsilon b_p z_p] \rho_p}{s_p e^{-ln4b_p}} \end{aligned} \quad (27)$$

where  $z = [z_p^w, z_{p+1}^w, \dots, z_N^w, z_p^x]$  is defined by connecting parameter size  $z_l^w, l \in \{p, p+1, \dots, N\}$  and activation size  $z_p^x$ . Below is the proof of the Eq.27.

##### Proof of Eq.27

In this section, we will prove the optimum of the proposed optimization problem. First, the optimization problem in Eq. 23 can be simplified as follows:

$$\begin{aligned} \min_{b, p} \quad & \xi \cdot \sum_{l=1}^p o(l) + \delta \cdot \sum_{l=p}^L o(l) + \epsilon \cdot (\sum_{l=p}^{L+1} b_l \cdot z_l) \\ \text{s.t.} \quad & \sum_{l=1}^L \frac{s_l e^{-ln4b_l}}{\rho_l} - \Delta \leq 0 \end{aligned} \quad (28)$$

We denote the objective function as  $f(\cdot)$ . The following Hessian matrix can check the convexity of the objective function:

$$\begin{aligned}
H &= \begin{bmatrix} \frac{\partial^2 f}{\partial b_p^2} & \frac{\partial^2 f}{\partial b_p \partial b_{p+1}} & \cdots & \frac{\partial^2 f}{\partial b_p \partial b_{L+1}} & \frac{\partial^2 f}{\partial b_p \partial p} \\ \frac{\partial^2 f}{\partial b_{p+1} \partial b_p} & \frac{\partial^2 f}{\partial b_{p+1}^2} & \cdots & \frac{\partial^2 f}{\partial b_{p+1} \partial b_{L+1}} & \frac{\partial^2 f}{\partial b_{p+1} \partial p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial b_{L+1} \partial b_p} & \frac{\partial^2 f}{\partial b_{L+1} \partial b_{p+1}} & \cdots & \frac{\partial^2 f}{\partial b_{L+1}^2} & \frac{\partial^2 f}{\partial b_{L+1} \partial p} \\ \frac{\partial^2 f}{\partial p \partial b_p} & \frac{\partial^2 f}{\partial p \partial b_{p+1}} & \cdots & \frac{\partial^2 f}{\partial p \partial b_{L+1}} & \frac{\partial^2 f}{\partial p^2} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ \epsilon z_p & 0 & \cdots & 0 & 0 \end{bmatrix} \succeq 0
\end{aligned} \tag{29}$$

We convert the simplified optimization problem in Eq.28 to Lagrange function below:

$$\begin{aligned}
\min_{\mathbf{b}, p} \quad & \xi \cdot \sum_{l=1}^p o(l) + \delta \cdot \sum_{l=p}^L o(l) + \epsilon \cdot (\sum_{l=p}^{L+1} b_l \cdot z_l) \\
& + \lambda \cdot (\sum_{l=1}^L \frac{s_l e^{-\ln 4 b_l}}{\rho_l} - \Delta)
\end{aligned} \tag{30}$$

We can solve the gradient vector of function  $f(\cdot)$  as follows:

$$\nabla f = \left[ \frac{\partial f}{\partial b_p} \frac{\partial f}{\partial b_{p+1}} \cdots \frac{\partial f}{\partial b_{N+1}} \frac{\partial f}{\partial p} \right] \tag{31}$$

where partial derivative of function  $f(\cdot)$  w.r.t.  $\mathbf{b}$  and  $p$  can be expressed as:

$$\frac{\partial f}{\partial b_l} = z_l, \quad l \in \{p, p+1, \dots, N+1\} \tag{32}$$

$$\frac{\partial f}{\partial p} = \xi o(p) - \delta o(p) - \epsilon b_p z_p \tag{33}$$

We denote the constrain fuction as  $g(\cdot)$ , which can be expressed as:

$$g(\mathbf{b}, p) = \sum_{l=1}^L \frac{s_l e^{-\ln 4 b_l}}{\rho_l} - \Delta \tag{34}$$

Similarly, we can derive gradient vector of function  $g(\mathbf{b}, p)$  in Eq. 35

$$\nabla g = \left[ \frac{\partial g}{\partial b_p} \frac{\partial g}{\partial b_{p+1}} \cdots \frac{\partial g}{\partial b_{N+1}} \frac{\partial g}{\partial p} \right] \tag{35}$$

where partial derivatives can be detailed as follows:

$$\frac{\partial g}{\partial b_l} = -\frac{\ln 4 \cdot s_l}{\rho_l} e^{-\ln 4 \cdot b_l}, \quad l \in \{p, p+1, \dots, N+1\} \tag{36}$$

$$\frac{\partial g}{\partial p} = -\frac{s_p \cdot e^{-\ln 4 b_p}}{\rho_p} \tag{37}$$

By using KKT conditions on the Lagrange function defined in Eq.30, we have the following equations:

$$\begin{aligned}
z_p &= \lambda \cdot \frac{\ln 4 \cdot s_p}{\rho_p} e^{-\ln 4 b_p} \\
z_{p+1} &= \lambda \cdot \frac{\ln 4 \cdot s_{p+1}}{\rho_{p+1}} e^{-\ln 4 b_{p+1}} \\
&\vdots \\
z_{N+1} &= \lambda \cdot \frac{\ln 4 \cdot s_{N+1}}{\rho_p} e^{-\ln 4 b_{N+1}} \\
\xi o(p) - \delta o(p) - \epsilon b_p z_p &= \lambda \cdot \frac{s_p \cdot e^{-\ln 4 b_p}}{\rho_p}
\end{aligned} \tag{38}$$

Solving these equations, we have:

$$\begin{aligned}
\frac{z_p \rho_p}{s_p e^{-\ln 4 b_p}} &= \frac{z_{p+1} \rho_{p+1}}{s_{p+1} e^{-\ln 4 b_{p+1}}} = \cdots = \frac{z_{N+1} \rho_{N+1}}{s_{N+1} e^{-\ln 4 b_{N+1}}} \\
&= \frac{\ln 4 \cdot [\xi o(p) - \delta o(p) - \epsilon b_p z_p] \rho_p}{s_p e^{-\ln 4 b_p}}
\end{aligned} \tag{39}$$

By further combining the first term and the last term in Eq.39 for resolution, we have:

$$b_p = \frac{\xi o(p) - \delta o(p) - \frac{z_p}{\ln 4}}{\epsilon z_p} \tag{40}$$

which means we can directly get quantization bit-width for layer  $p$  with given partition point  $p$ . In the rest of this section, we elaborate on the calculation of  $\mathbf{b}$  and  $p$  by proposing two algorithms according to the solution above.

From the solution in Eq.39 and Eq.40, we can observe that layer-wise quantization bit-width is solved by iteratively calculating the adversarial noise  $\bar{\sigma}^*$  and robustness parameter  $\rho_l$  from the first layer to the last layer of the first model segment. Multiple forwarding operations are involved in this process of calculating adversarial noise, which renders it impractical for real-time deployment. This is a common challenge in many optimization problems, particularly in systems with strict real-time constraints. To address this, many works in the literature resort to approximate or heuristic-based methods that trade off some degree of optimality for practical applicability. In this case, we design the inference serving strategy *QPART* with two following algorithms.

- 1) **Offline Quantization Algorithm:** The optimal layer-wise quantization bit-width is solved by the offline quantization algorithm and it's pre-computed under various system conditions. By optimizing the layer-wise quantization offline, most of the calculation overhead can be dealt in advance and the inference serving system can hence respond to the requests in real-time.
- 2) **Online Serving Algorithm:** With quantization patterns been computed by offline algorithms, the online serving algorithm responds to the inference request by searching the layer-wise quantization pattern and partition point among pre-calculated results.

#### A. Offline Quantization Algorithm

The offline quantization algorithm is shown in Algorithm.1, which takes a parameterized model  $\theta$  as input, and generates

a set of model variants for all possible partition point and several different accuracy level. Specifically, the proposed *QPART* enumerate the partition point from the first layer to the last layer  $p \in \{1, 2, \dots, L\}$ , with 5 different accuracy degradation requirements  $a \in \{a_1, a_2, a_3, a_4, a_5\}$ , and optimize the quantization bit-width  $b_a^p$  accordingly. Then the set of quantization pattern  $\{(\mathbf{b}_a^p, p)\}_\theta$  will be stored for the online serving algorithm in runtime.

---

**Algorithm 1** Offline Model Quantization Algorithm for Inference offloading

---

**Require:** Model  $\theta$

**Ensure:** Set of quantized and partitioned model variants  $\{\theta_a^p\}$ , where  $a \in \{a_1, a_2, a_3, a_4, a_5\}, p \in \{1, 2, \dots, L\}$

```

1: for  $a$  in  $\{a_1, a_2, a_3, a_4, a_5\}$  do
2:   for  $p = 1, L$  do
3:     for  $l = p, 1$  do
4:       if  $l == p$  then
5:         Calculate bit-width at partition point  $p$ :  $b_p = \frac{\xi o(p) - \delta o(p) - \frac{z_i}{l n 4}}{\epsilon z_p}$ 
6:       end if
7:       Calculate average adversarial noise  $\bar{\sigma}^*$  with given dataset  $\mathcal{D}$ 
8:       Incrementally introduce noise into the parameter of layer  $l$ , observe the accuracy degradation, and record the noise threshold  $\sigma_l$  at which the accuracy degradation equals  $a$ .
9:       Calculate robustness parameter  $\rho_l$  by Eq.20
10:      Calculate  $s_l$  using Eq. 18, by fixing the  $b_l$ 
11:      Calculate the quantization bit width  $b_l$  for layer  $l$ , by Eq. 27
12:    end for
13:    Save  $\mathbf{b}_a^p = (b_1, b_2, \dots, b_L)$ 
14:  end for
15: end for
16: return Set of optimized quantization and partition pattern  $\{(\mathbf{b}_a^p, p)\}_\theta$ 

```

---

For each model quantization pattern  $\mathbf{b}_a^p$  with given partition point  $p$  and accuracy degradation requirement  $a$ , to optimize the layer-wise quantization bit width, we start from the layer  $p$ . The bit width for the layer at the partition point can be directly derived from Eq.27. Subsequently, the bit width for the preceding layer  $b_l, l \in p-1, p-2, \dots, 1$ , is solved sequentially from layer  $p-1$  to layer 1 by substituting the parameters of the previous-calculated layer  $b_{l+1}, \rho_{l+1}, s_{l+1}, z_{l+1}$ , and current layer  $\rho_l, s_l, z_l$  with actual values in Eq.27.

### B. Online Serving Algorithm

In the online serving stage, the *QPART* responds to the request from the edge devices by searching for the optimal quantization and partitioning pattern  $(\mathbf{b}_{a^*}^p, p^*)_\theta$  within the set  $\{(\mathbf{b}_a^p, p)\}_\theta$ , that meets the requirement of the accuracy degradation, and then quantizing the pre-trained model  $\theta$ , shown in Algorithm.2

---

**Algorithm 2** Online Inference Serving Algorithm for Edge Devices

---

**Require:** Inference Request from Edge Device  $(\theta, a, r, \pi, \gamma_{local}, f_{local}, \kappa)$

**Ensure:** Optimized Model Variant  $\theta^*$

```

1: Select  $a^*$  to be the maximum from set  $\{a_1, a_2, a_3, a_4, a_5\}$  without exceeding  $a$ 
2: for  $p = 1 : L$  do
3:   Calculate Objective Function Value in Eq.17, denoted as  $\hat{j}_p$ 
4: end for
5: Select  $p^*$  to be the value from  $\{1, 2, \dots, L\}$  that minimizes the objective function value
6: Load optimized quantization and partition pattern  $(\mathbf{b}_{a^*}^{p^*}, p^*)_\theta$  from  $\{(\mathbf{b}_a^p, p)\}_\theta$ 
7: Quantize the model  $\theta$  to be  $\theta^*$  according to Eq.10
8: return  $\theta^*$ 

```

---

## V. SIMULATIONS AND NUMERICAL RESULTS

In this section, we present the numerical results obtained from our simulation platform, which validate the effectiveness of our proposed model quantization and partitioning algorithm for inference offloading. Specifically, we have implemented a DNN-based classifier for the MNIST handwritten digits dataset [59] to demonstrate the inference offloading capabilities of our approach. This DNN comprises six fully connected layers, as illustrated in Fig.4. We have meticulously gathered and analyzed the simulation results, including time and energy consumption measurements, overall costs, and accuracy. Furthermore, to gain a more comprehensive perspective on compression-based inference offloading schemes and thoroughly examine the efficacy of our proposed algorithm *QPART* in the context of compression-based inference offloading, we have performed comparative tests with auto-encoder-based offloading [35], model-pruning-based offloading [44] [45], and direct offloading without the application of any optimization methods.

We set up the simulator with Python script, within which there are two components: executing module, communication module and performance module. The executing module simulates the processing behaviors of edge devices and server according to the parameters shown in Table II, and execute the two model segments with different processing profile. The communication module simulates the transmission of the model parameter and intermediate output between the edge device and the server based on the wireless communication model defined in Section III. The performance module collects the performance metrics from the other two modules, such as time consumptions and energy consumptions. Specific simulation parameters are summarized in Table II:

First we enumerate the model partition point and calculate the quantization bit-width for subsequent layers. As the partition point is incrementally shifted from the first layer towards the final layer, we plot the total time consumption, energy consumption and server cost, which is shown in Fig.5.

Analysis of Fig.5 reveals a trend: As the partition point moves closer to the last layer, the time consumption and power



TABLE II  
SIMULATION SETTINGS

Parameters	Interpretation	Value
$\gamma_{local}$	Average clock cycles needed for a MAC Operation on mobile device	5
$\gamma_{server}$	Average clock cycles needed for a MAC Operation on server	5/4
$f_{local}$	Mobile device clock rate	200 MHz
$f_{server}$	Server clock rate	3 GHz
$\pi_n$	Mobile device transmit power	1 watt
$r_{n,m}$	Channel capacity	200 Mbps
$\omega$	Time consumption weight	1
$\tau$	Energy consumption weight	1
$\kappa_n$	Energy efficiency parameter of mobile device	$3 \times 10^{-27}$
$\eta_m$	Energy efficiency parameter of edge server	$3.75 \times 10^{-27}$

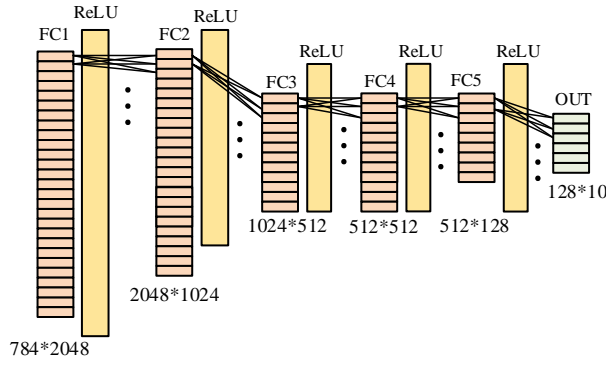


Fig. 4. Inference Task Model

consumption increase while the server cost decreases. The reason for the increase in time consumption and energy consumption along the partition point moving towards the output layer is that the more model parameters would be counted into the communication payload from the *QPART* platform to the device. Meanwhile the server cost is reducing because the shared computation workload on the server decrease while more calculations done in the device. Therefore, a trade-off in computation workload sharing becomes evident: attempting to save time and energy by allocate more computation on the server increase server cost. Specifically, we compared the layer-wise performance metrics under two strategies, the *QPART* and the inference service without optimization, which means the model segments are transmitted from server to the devices without initial pre-trained model intact. From Fig.5, we can find out the *QPART* reduce the power consumption and time consumption significantly in all possible partition layers.

By quantizing the model parameter properly, the time consumption and power consumption are hugely reduced, however the lower bit width doesn't come for free. According to Eq.18 and Eq.19, the quantization noise will increase while the quantization bit width decrease, which will degradate the inference accuracy. Fig.6 illustrates this trade-off between accuracy and model parameter size by establishing various

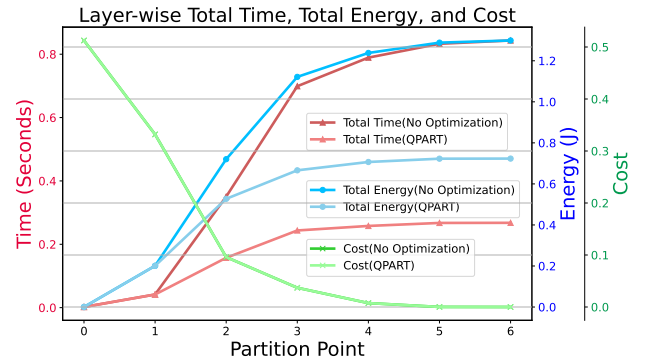


Fig. 5. Layer-wise Performance Comparison

accuracy degradation thresholds and testing the optimized parameter size. This relationship is discerned by aggregating the sizes of the parameters across all layers. It is evident from the figure that the overall model parameter size can decrease exponentially as the required accuracy level is lowered.

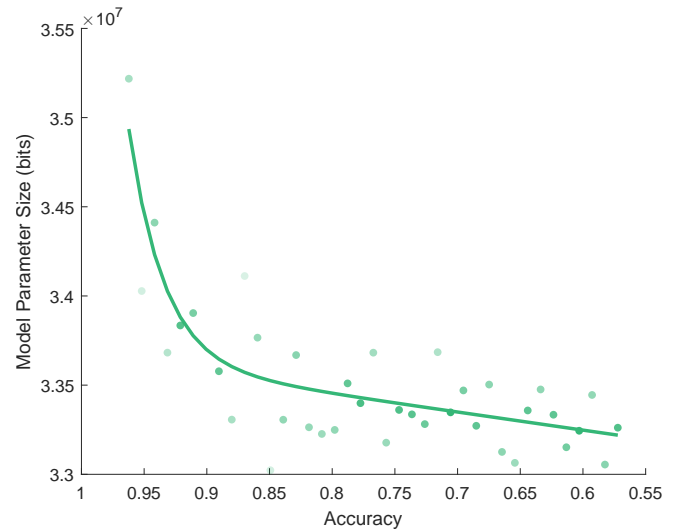


Fig. 6. Optimized Model Size v.s. Accuracy

Fig.7, Fig.8, and Fig.9 compare the proposed algorithm

*QPART* with autoencoder-based [35] and model pruning-based [44] [45] inference task offloading algorithms. The auto-encoder-based algorithm compresses the intermediate activation by inserting an encoder and decoder on both the device and the server sides. The model pruning-based method trims neurons in the layers designated for offloading to the server, ensuring that the number of pruned neurons is balanced to maintain accuracy degradation comparable to that of the *QPART* algorithm. This strategic pruning of neurons facilitates a reduction in the size of the communication payload. Fig.7 illustrates the objective function values of four offloading schemes, varying according to the partition point. The *QPART* algorithm demonstrates the ability to achieve the lowest overall cost. Notably, the autoencoder-based approach tends to incur a significantly higher overall cost than the other three schemes, a finding further substantiated by the subsequent figures.

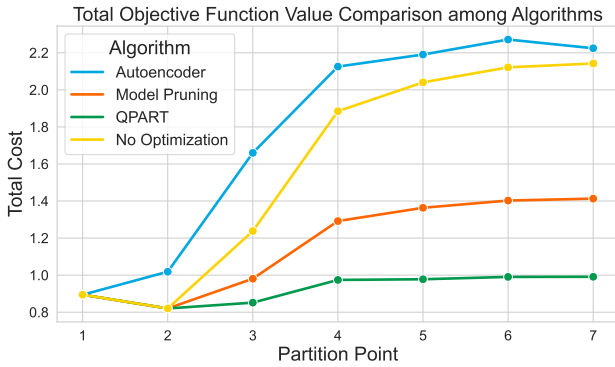


Fig. 7. Layer-wise Total Cost Comparison

Fig.8 and Fig.9 provide detailed breakdowns of the time and energy consumption for the four offloading schemes. It is readily apparent that the *QPART* algorithm consistently exhibits the lowest energy and time consumption, regardless of the partition point. Due to the incorporation of additional encoding and decoding layers, the auto-encoder-based method incurs greater time consumption and, consequently, higher energy consumption than the other three schemes. Despite this, it slightly reduces communication payload, as shown in Fig.10.

Table III presents the inference accuracy associated with the above four offloading schemes. We evaluated the model's accuracy using the MNIST test dataset, applying four distinct offloading strategies. In scenarios where no optimization algorithm is incorporated into the offloading scheme, the inference accuracy remains consistent with that of the unmodified neural network. The auto-encoder-based method tends to demonstrate improved prediction accuracy when more layers are processed locally. This trend can be attributed to the layers closer to the input being more susceptible to the recovery noise introduced by the autoencoder. The proposed *QPART* algorithm maintains accuracy successfully and surpasses the other two compression-based methods in performance.

We deployed a convolutional neural network (CNN) on the SVHN, CIFAR10, CIFAR100, and ImageNet datasets to validate our proposed framework across a broader range of datasets and baseline neural networks. In addition, we imple-

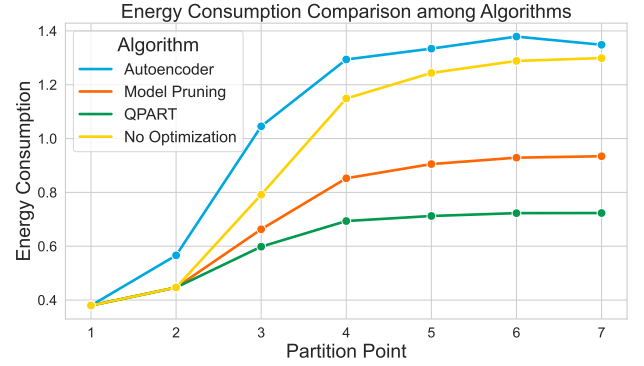


Fig. 8. Layer-wise Energy Consumption Comparison

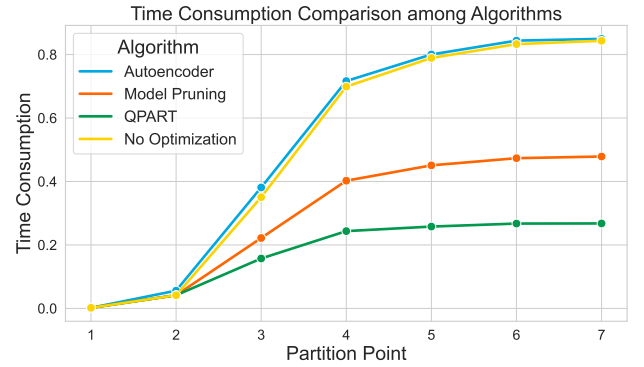


Fig. 9. Layer-wise Time consumption Comparison

mented ResNet architectures, including ResNet18, ResNet34, and ResNet50, on the ImageNet dataset to further benchmark our framework. Table IV illustrates the efficacy of our proposed *QPART* when applied across these datasets and models. Notably, we observed a significant reduction in communication payload, calculated by dividing the optimized parameter size by the initial parameter size, ranging from 11.88% to 18.12%, with negligible accuracy degradation, ranging from 0.08% to 0.66%. These results underscore the versatility and efficiency of our proposed model.

## VI. CONCLUSION

In this paper, we have proposed *QPART*, a neural network inference serving system, emphasizing the development of a flexible, accuracy-aware, and retraining-free tailored for different edge devices. This was achieved by integrating model quantization and model partitioning. We have modeled the time and energy consumption associated with local inference, data transmission, and remote inference. Additionally, our approach includes an accuracy degradation measurement within the joint optimization problem, allowing us to identify the optimal quantization bit-width and partition point while considering accuracy requirements. Our simulation results demonstrate a significant reduction in latency and energy consumption, underscoring the effectiveness of our proposed *QPART* algorithm. Looking ahead, our future work will explore global scheduling and allocation strategies. We aim to

TABLE III  
ACCURACY OF ALGORITHMS AT DIFFERENT PARTITION POINTS

Partition Point	Auto-Encoder (%)	No Optimization (%)	Model Pruning (%)	<i>QPART</i> (%)
0	93.63%	96.19%	95.03%	96.10%
1	93.64%	96.19%	95.03%	96.21%
2	95.91%	96.19%	95.03%	96.21%
3	96.27%	96.19%	95.03%	96.21%
4	95.99%	96.19%	95.03%	96.18%
5	96.15%	96.19%	95.03%	96.20%

TABLE IV  
ACCURACY ON BASELINE MODELS AND DATASETS

	SVHN	CIFAR10	CIFAR100	Resnet18 (ImageNet)	Resnet34 (ImageNet)
Initial Parameter Size (MB)	5.89	46.55	46.9	107.96	180.01
Optimized Parameter Size (MB)	0.79	5.53	6.35	19.00	32.62
Communication Payload Compression Ratio	13.45%	11.88%	13.53%	17.60%	18.12%
Initial Accuracy	86.88%	89.52%	67.35%	62.75%	67.21%
Accuracy Optimized by AOIFE	86.79%	89.4%	67.12%	62.09%	66.63%
Accuracy Degradation	0.08%	0.12%	0.23%	0.66%	0.58%

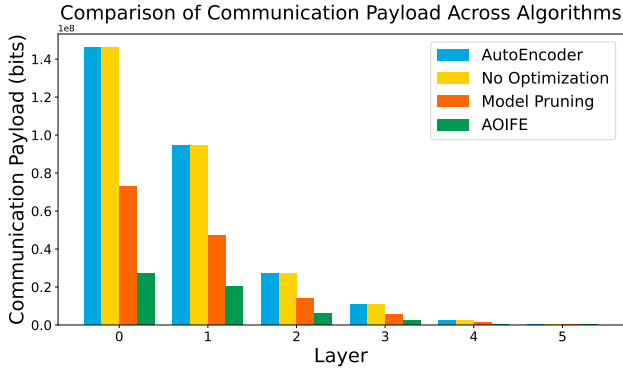


Fig. 10. Layer-wise Communication Payload Comparison

propose a comprehensive hierarchical optimization framework for inference offloading in mobile edge computing systems, which we anticipate will further enhance the system's overall efficiency and effectiveness in managing complex computational tasks.

#### ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under Grants Nos 2315851 and 2106634, a Sony Faculty Innovation Award (Contract AG3ZURVF) and a Cisco Research Award (Contract 878201).

#### REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.
- [6] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [7] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.
- [8] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [9] S. Hosseininoorbin, S. Layeghy, M. Sarhan, R. Jurdak, and M. Portmann, "Exploring edge tpu for network intrusion detection in iot," *Journal of Parallel and Distributed Computing*, vol. 179, p. 104712, 2023.
- [10] S. Mittal, "A survey on optimized implementation of deep learning models on the nvidia jetson platform," *Journal of Systems Architecture*, vol. 97, pp. 428–442, 2019.
- [11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [12] M. Tan, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.
- [13] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 2002, pp. 55–69.
- [14] A. Bakhtiarnia, Q. Zhang, and A. Iosifidis, "Single-layer vision transformers for more accurate early exits with less overhead," *Neural Networks*, vol. 153, pp. 461–473, 2022.
- [15] M. Lin, Y. Zhang, Y. Li, B. Chen, F. Chao, M. Wang, S. Li, Y. Tian, and R. Ji, "1xn pattern for pruning convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 3999–4008, 2022.
- [16] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *arXiv preprint arXiv:1810.05270*, 2018.
- [17] D. Hu and B. Krishnamachari, "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)*. IEEE, 2020, pp. 157–163.
- [18] H. Chen, L. Zeng, S. Yu, and X. Chen, "Knowledge distillation for mobile edge computation offloading," *arXiv preprint arXiv:2004.04366*, 2020.
- [19] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 1, pp. 615–629, 2017.
- [20] L. Liu, H. Li, and M. Gruteser, "Edge assisted real-time object detec-

- tion for mobile augmented reality,” in *The 25th annual international conference on mobile computing and networking*, 2019, pp. 1–16.
- [21] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia *et al.*, “Machine learning at facebook: Understanding inference at the edge,” in *2019 IEEE international symposium on high performance computer architecture (HPCA)*. IEEE, 2019, pp. 331–344.
  - [22] S. Bhattacharya and N. D. Lane, “From smart to deep: Robust activity recognition on smartwatches using deep learning,” in *2016 IEEE International conference on pervasive computing and communication workshops (PerCom Workshops)*. IEEE, 2016, pp. 1–6.
  - [23] Q. Zhou, Z. Qu, S. Guo, B. Luo, J. Guo, Z. Xu, and R. Akerkar, “On-device learning systems for edge intelligence: A software and hardware synergy perspective,” *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11916–11934, 2021.
  - [24] M. M. Amiri and D. Gündüz, “Federated learning over wireless fading channels,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.
  - [25] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
  - [26] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.
  - [27] I. Al-Omari, S. Hadayeghparast, and H. Karimipour, “Application of deep learning on iot-enabled smart grid monitoring,” *AI-Enabled Threat Detection and Security Analysis for Industrial IoT*, pp. 77–103, 2021.
  - [28] B. Hallinan and T. Striphas, “Recommended for you: The netflix prize and the production of algorithmic culture,” *New media & society*, vol. 18, no. 1, pp. 117–137, 2016.
  - [29] M. S. Basarlsan, F. Kayaalp *et al.*, “Sentiment analysis with machine learning methods on social media,” *Advances in Distributed Computing and Artificial Intelligence Journal*, 2020.
  - [30] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” *arXiv preprint arXiv:1510.00149*, 2015.
  - [31] X. Jiao, V. Akhlaghi, Y. Jiang, and R. K. Gupta, “Energy-efficient neural networks using approximate computation reuse,” in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1223–1228.
  - [32] E. Li, L. Zeng, Z. Zhou, and X. Chen, “Edge ai: On-demand accelerating deep neural network inference via edge computing,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 447–457, 2019.
  - [33] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, and P. Frossard, “Adaptive quantization for deep neural network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
  - [34] Z. Xu, L. Zhao, W. Liang, O. F. Rana, P. Zhou, Q. Xia, W. Xu, and G. Wu, “Energy-aware inference offloading for dnn-driven applications in mobile edge clouds,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 4, pp. 799–814, 2020.
  - [35] S. Yao, J. Li, D. Liu, T. Wang, S. Liu, H. Shao, and T. Abdelzaher, “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.
  - [36] T. Mohammed, C. Joe-Wong, R. Babbar, and M. Di Francesco, “Distributed inference acceleration with adaptive dnn partitioning and offloading,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 854–863.
  - [37] Q. Yang, X. Luo, P. Li, T. Miyazaki, and X. Wang, “Computation offloading for fast cnn inference in edge computing,” in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 2019, pp. 101–106.
  - [38] H. B. Beytur, A. G. Aydin, G. de Veciana, and H. Vikalo, “Optimization of offloading policies for accuracy-delay tradeoffs in hierarchical inference,” in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 1989–1998.
  - [39] A. Ben Sada, A. Khelloufi, A. Naouri, H. Ning, N. Aung, and S. Dhelim, “Multi-agent deep reinforcement learning-based inference task scheduling and offloading for maximum inference accuracy under time and energy constraints,” *Electronics*, vol. 13, no. 13, p. 2580, 2024.
  - [40] F. Romero, Q. Li, N. J. Yadwadkar, and C. Kozyrakis, “{INFaaS}: Automated model-less inference serving,” in *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, 2021, pp. 397–411.
  - [41] X. Wang, M. Magno, L. Cavigelli, and L. Benini, “Fann-on-mcu: An open-source toolkit for energy-efficient neural network inference at the edge of the internet of things,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4403–4417, 2020.
  - [42] I. Chakraborty, D. Roy, I. Garg, A. Ankit, and K. Roy, “Constructing energy-efficient mixed-precision neural networks through principal component analysis for edge intelligence,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 43–55, 2020.
  - [43] M. S. Barbosa, R. G. Pacheco, R. S. Couto, D. S. Medeiros, and M. E. M. Campista, “Decision early-exit: An efficient approach to hasten offloading in branchynets,” in *2022 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE, 2022, pp. 1–6.
  - [44] J. Ren, D. Yang, Y. Yuan, H. Wei, and Z. Wang, “A deep reinforcement learning based research for optimal offloading decision,” *AIP Advances*, vol. 13, no. 8, 2023.
  - [45] W. Shi, Y. Hou, S. Zhou, Z. Niu, Y. Zhang, and L. Geng, “Improving device-edge cooperative inference of deep learning via 2-step pruning,” in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 1–6.
  - [46] N. Li, A. Iosifidis, and Q. Zhang, “Graph reinforcement learning-based cnn inference offloading in dynamic edge computing,” 2022.
  - [47] R. G. Pacheco, M. Shifrin, R. S. Couto, D. S. Menasché, M. K. Hanawal, and M. E. M. Campista, “Adaee: Adaptive early-exit dnn inference through multi-armed bandits,” in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 3726–3731.
  - [48] Y. Duan and J. Wu, “Optimizing job offloading schedule for collaborative dnn inference,” *IEEE Transactions on Mobile Computing*, 2023.
  - [49] X. Zhao, M. Hosseinzadeh, N. Hudson, H. Khamfroush, and D. E. Lucani, “Improving the accuracy-latency trade-off of edge-cloud computation offloading for deep learning services,” in *2020 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2020, pp. 1–6.
  - [50] A. B. Sada, A. Khelloufi, A. Naouri, H. Ning, and S. Dhelim, “Selective task offloading for maximum inference accuracy and energy efficient real-time iot sensing systems,” *arXiv preprint arXiv:2402.16904*, 2024.
  - [51] S. Lahmer, A. Khoshshirat, M. Rossi, and A. Zanella, “Energy consumption of neural networks on nvidia edge boards: an empirical model,” in *2022 20th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt)*. IEEE, 2022, pp. 365–371.
  - [52] K. De Vogelee, G. Memmi, P. Jouvelot, and F. Coelho, “The energy/frequency convexity rule: Modeling and experimental validation on mobile devices,” in *Parallel Processing and Applied Mathematics: 10th International Conference, PPAM 2013, Warsaw, Poland, September 8-11, 2013, Revised Selected Papers, Part I 10*. Springer, 2014, pp. 793–803.
  - [53] T. D. Burd and R. W. Brodersen, “Processor design for portable systems,” *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 13, no. 2-3, pp. 203–221, 1996.
  - [54] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song *et al.*, “Going deeper with embedded fpga platform for convolutional neural network,” in *Proceedings of the 2016 ACM/SIGDA international symposium on field-programmable gate arrays*, 2016, pp. 26–35.
  - [55] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, “A survey of quantization methods for efficient neural network inference,” *arXiv preprint arXiv:2103.13630*, 2021.
  - [56] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized neural networks: Training neural networks with low precision weights and activations,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.
  - [57] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, “Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients,” *arXiv preprint arXiv:1606.06160*, 2016.
  - [58] X. Li, J. Chen, X. Ling, and T. Wu, “Deep reinforcement learning-based anti-jamming algorithm using dual action network,” *IEEE Transactions on Wireless Communications*, 2022.
  - [59] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.