PhD Forum Abstract: Multi-View Service Provisioning in Cloud-Edge-End Networks with Hierarchical Resources

Haodong Zou
Hong Kong Baptist University, Hong Kong and BNU-HKBU United International College, Zhuhai, China
haodongzou@uic.edu.cn

ABSTRACT

With the surge of end devices and intelligent services, computing resource has begun to migrate from the cloud to end devices to meet the growing demand of users, forming a hierarchical resource distribution pattern in cloud-edge-end networks. Existing research work focuses on using edge computing technique to build a cloud-edge collaborative service offloading and task scheduling method to reduce service delay or energy consumption. However, different user groups and application scenarios may have different preferences for service quality requirements even for the same kind of service. For example, video analysis in autonomous driving focuses more on delay while video analysis in surveillance focuses more on accuracy. Meanwhile, heterogeneous cloud-edge-end devices have significant differences in the amount of resources, which poses great challenges for efficient provisioning of services. To solve this problem, we intend to propose multi-view service provisioning method in cloud-edge-end networks with hierarchically distributed resources. Firstly, we design a mapping scheme between service quality and heterogeneous resource occupation to estimate the amount of resources required for a given requirement. Secondly, we utilize model compression methods to customize powerful large models into smaller and lighter one according to the requirements of tasks. Thirdly, as resources are distributed hierarchically in cloud-edge-end networks, efficient service placement should be carried out with the goal of achieving diverse needs. The effectiveness of the proposed method is demonstrated through numerical simulations compared to state-of-the-art baselines and experiments on an implemented prototype system.

KEYWORDS

Service Provisioning, Multi-View Requirements, Heterogeneous Resources, Cloud-Edge-End Networks

1 PROBLEM DESCRIPTION

In this work, we study the service provisioning problem in a cloud-edge-end networks, where the resources, such as storage, computing, and bandwidth, are hierarchically distributed among these three main levels. As depicted in Figure 1, a cloud server may equipped with rich resources while edge servers and end devices, such as smart phone, personal computer, and augmented reality glasses, usually are resource-constrained, forming a heterogeneous connected networks through wired and wireless manners.

The service provisioning problem is to provide satisfying Quality of Service (QoS) to the end devices by selectively placing on-demand services in accordance with the available amount of resource in cloud-edge-end networks. For each service, the end devices may have distinct orientations on the QoS, e.g., accuracy, delay, and energy consumption. Simply placing one version of certain service can not guarantee fulfillment of diverse requirements from different end devices. Besides, resources in cloud-edge-end networks are hierarchically distributed, which implies that identical placement of service at cloud, edge, and end devices may not be possible. Last but not least, services placed on certain device may encounter with unexpected large amount of tasks, affecting the service delay if there is a backlog of tasks. How to utilize the potential of interconnected devices in cloud-edge-end networks to further mitigate the QoS reduction in busy traffic scenarios remains unsolved.

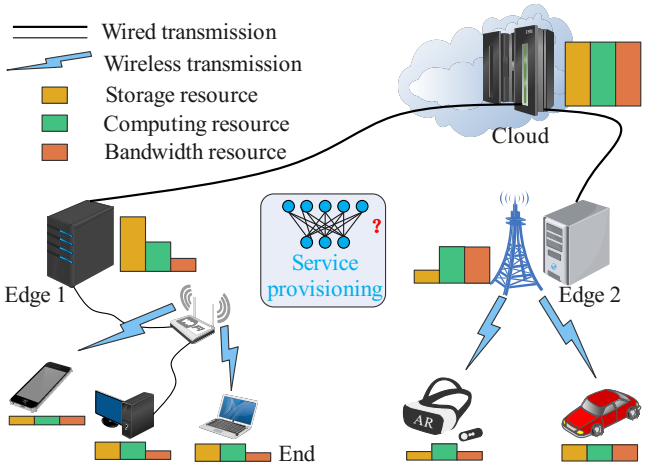

Figure 1: The studied cloud-edge-end networks. The resources are hierarchically distributed among cloud, edge, and end devices.

2 RELATED WORKS

Many efforts are devoted to the research of service provisioning.

2.1 Model Compression and Distillation

Liu *et al.* [5] adopted a simple stochastic structure sampling scheme for the training of a pruning network. And an evolutionary procedure is employed to search for pruned networks with good performance. Elsen *et al.* [1] further expanded the efficient building blocks for neural network. They advocated for the replacement of these dense primitives with their sparse counterparts. Wang *et al.* [7] investigated compression and efficient inference methods for large language models (LLMs) by incorporating the fine-tuning cost and generalization characters.

2.2 Service Placement and Task Scheduling

Ning *et al.* [6] proposed to solve the dynamic service placement problem with edge storage and service execution delay constraints.

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00067

Li *et al.* [4] designed a collaborative service placement and task scheduling framework. Feng *et al.* [2] proposed a novel spatial-temporal collaboration of service placement and task scheduling to minimize the overall cost of edge systems. Li *et al.* [3] proposed an online algorithm based on the two-timescale Lyapunov optimization in a stochastic network environment.

3 MULTI-VIEW SERVICE PROVISIONING

We intend to solve service provisioning problem in three steps.

3.1 Multi-View Model Compression

The first step is to generate a series of models with preference from a large and accurate models. In Figure 2, a neural network model based service can be pruned into a smaller model with less intermediate layer neurons, thus reducing the required computation and achieving low delay of service. The low delay service is suitable for placement on edge due to their close proximity to end devices. On top of the neural network pruning, model quantization can also be applied to scale down the model size as shown in Figure 2. With smaller model size, the energy consumption of the subsequent model placement and model inference when providing service is expected to diminish. For those mobile end devices, such as smart phone and augmented reality glasses, the most important concern is their limited battery life. The energy-saving services may directly placed on the end devices to fulfill their diverse QoS.

3.2 Hierarchical Service Placement

After model compression, we design hierarchical service placement scheme in the second step. Recall that in Figure 1, the studied cloud-edge-end networks have imbalanced resource distribution. Aiming at an efficient and adaptive service placement, we intend to design a bidirectional mapping between QoS of certain model and the heterogeneous resource occupation. The involved model properties include parameter (weight) quantity, the data type (precision of number), and the complexity of related operations etc. These properties can be used to estimate how much memory, computation capacity, and bandwidth the device should supply to support the targeted QoS. Conversely, given a resources status on a device, we can utilize the mapping relation to obtain the maximal model properties the device can carry. In this way, tasks with different QoS orientations can be served with on-demand and personalized service in accordance with the hierarchical system resources.

3.3 Collaborative Task Scheduling

The services of multi-view QoS are adaptively placed into the cloud-edge-end networks and normally are able to satisfy end devices' requirements. In the rare situation, there will be large amount traffic of tasks, leading to a backlog of tasks on single device. This situation may cause long delay of services and harm the fulfillment of low delay QoS. As cloud-edge-end networks are an interconnected system, the instant peak requirements of tasks could be scheduled to idle peer devices for service provision. To achieve this, we design an edge-edge collaboration mechanism to relieve the large delay problem and improve the system resource utility. By reusing the associated QoS of placed services, we can efficiently filter out unsatisfying services and build a collection of feasible services.

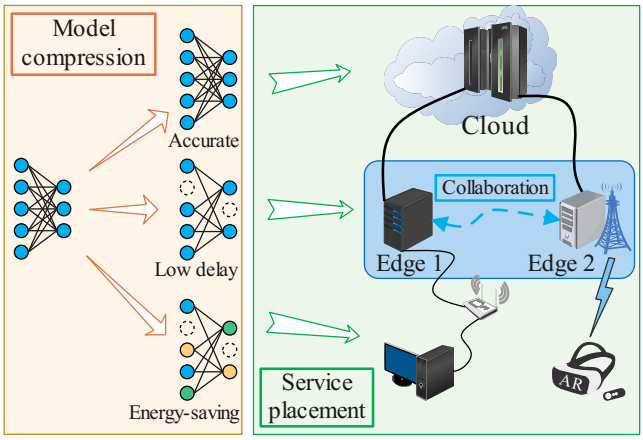

Figure 2: Multi-view service provisioning in hierarchical cloud-edge-end networks.

4 PRELIMINARY

The research on **collaborative task scheduling** has been studied previously, which is under reviewed by a conference currently.

5 FUTURE WORK

In the future, we would focus on the on-demand and resource-aware model compression research to explore the personalized service scaling. On the basis of model compression, we intend to further studied the flexible service provisioning in a dynamic cloud-edge-end environment. Combining three research together, a prototype system of multi-view service provisioning is expected to be designed and implemented on commercial available devices.

ACKNOWLEDGMENTS

I have been in the Ph.D. program for two years. My expected graduation date is September 2025. I want to thank Prof. Tian Wang (Supervisor) and Jianxiong Guo for their valuable advice.

REFERENCES

[1] Erich Elsen, Marat Dukhan, Trevor Gale, and Karen Simonyan. 2020. Fast sparse convnets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14629–14638.

[2] Chunhui Feng, Qinghai Yang, Tony QS Quek, Weihua Wu, and Kun Guo. 2023. Spatially-Temporally Collaborative Service Placement and Task Scheduling in MEC Networks. *IEEE Transactions on Vehicular Technology* (2023).

[3] Xin Li, Xinglin Zhang, and Tiansheng Huang. 2023. Joint task offloading and service placement for mobile edge computing: An online two-timescale approach. *IEEE Transactions on Cloud Computing* (2023).

[4] Yuqing Li, Wenkuan Dai, Xiaoying Gan, Haiming Jin, Luoyi Fu, Huadong Ma, and Xinbing Wang. 2021. Cooperative service placement and scheduling in edge clouds: A deadline-driven approach. *IEEE Transactions on Mobile Computing* 21, 10 (2021), 3519–3535.

[5] Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Kwang-Ting Cheng, and Jian Sun. 2019. MetaPruning: Meta Learning for Automatic Neural Network Channel Pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[6] Zhaolong Ning, Peiran Dong, Xiaojie Wang, Shupeng Wang, Xiping Hu, Song Guo, Tie Qiu, Bin Hu, and Ricky YK Kwok. 2020. Distributed and dynamic service placement in pervasive edge computing networks. *IEEE Transactions on Parallel and Distributed Systems* 32, 6 (2020), 1277–1292.

[7] Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. 2024. Model Compression and Efficient Inference for Large Language Models: A Survey. arXiv:2402.09748 [cs.CL]