PhD Forum Abstract:Exploring Service Placement and Request Scheduling Based on Cooperative Edge Computing in AIoT

Yuzhu Liang
Beijing Normal University
Beijing, China
cs_yuzhuliang@163.com

ABSTRACT

The rapid growth in data generated by the Artificial Internet of Things (AIoT) necessitates an increase in computational power and presents challenges to cloud infrastructure, including traffic congestion and latency issues in AIoT systems. This trend has fostered a shift toward edge-layer computation, with cooperative edge computing emerging as a potent solution to these challenges. However, the diversity and heterogeneity of AIoT systems present significant challenges with regard to service placement, cross-regional request scheduling, and efficient resource caching. My research aims to enhance cooperative edge computing by designing an optimized clustering algorithm for efficient service placement and data processing, developing a cooperative edge request scheduling method using digital twin technology to minimize system transmission delays, and devising a resource caching method employing deep reinforcement learning in cooperative game scenarios to optimize resource allocation. This research endeavors to enhance service placement and request scheduling efficiency, thereby offering substantial computational support to AIoT systems.

KEYWORDS

Cooperative edge computing, Service placement, Request scheduling, Deep reinforcement learning

1 EDGE COLLABORATION BASED SERVICE PLACEMENT

Leveraging cloud-edge collaboration, service placement has emerged as a pivotal trend in AIoT systems[2, 4]. These algorithms largely depend on placement strategies that prioritize areas of highest load and user density. However, this approach necessitates uploading significant volumes of data to cloud servers, resulting in issues such as prolonged transmission times and reduced efficiency in real-time data processing. Edge collaboration is recognized as a promising solution to address the efficient data processing problem[6, 9]. However, two significant challenges remain to be resolved. The first challenge involves the placement of edge services; improper placement may lead to increased transmission delays and unbalanced loads. The second challenge pertains to leveraging collaborative edge servers to enhance accuracy despite the storage, communication, computation, and load constraints inherent in edge servers.

To address these challenges, we introduce a novel distributed collaborative edge framework, comprising two key stages: service placement and efficient data processing [5]. We propose an optimized clustering algorithm for the former, which considers the data distribution on the IoT layer and the constraints of the edge layer. The algorithm exhibits an approximate performance complexity of $O(\ln m)$, where m represents the number of users' tasks. For the latter, we introduce an efficient gossip-based data processing algorithm, by combining model parameters among edge servers, that fully utilizes edge collaboration to enhance efficient data processing. Both theoretical analysis and experimental results reveal that our algorithm achieves an average improvement in data processing accuracy of 9.02% and a reduction in delay of 36.61%, surpassing the performance of state-of-the-art works in various scenarios. We highlight that this framework is poised to play an increasingly important role in ensuring high QoS for AIoT systems.

2 DT-BASED CROSS-REGIONAL REQUEST SCHEDULING

After service placement in a distributed edge computing environment, individual edge nodes (ENs) often lack awareness of the global distribution of resources, which poses challenges to cross-network/region collaborative request scheduling[11, 12]. The integration of Digital Twins (DTs) with edge computing presents a viable solution to bridge this coordination gap. By constructing a DT model that reflects the status and availability of resources at each EN, it becomes possible to simulate the global resource landscape within the edge computing environment. This approach not only aids in the accurate mapping of resources but also facilitates more informed decision-making regarding resource allocation and service deployment[3].

Addressing the challenge of building an effective DT model necessitates a federated approach, taking into account the real-time statuses of ENs, particularly their computing and storage capacities. Given the limited capacities of these ENs, a cooperative methodology is proposed for the initial construction of the DT model. Through this federated process, ENs collaborate by sharing their computational resources to develop a basic DT model that evolves over time into a more comprehensive global model. This evolution is facilitated by the exchange of model parameters instead of the original data, significantly reducing the latency associated with data transmission and enhancing the overall efficiency of the process[13].

Furthermore, the DT model plays a crucial role in identifying the specific locations of ENs where resources are located. Given the potential for multiple routing paths between directly connected ENs and the target EN housing the desired resources, the identification of the most efficient path is essential. Longer paths involve additional hops, which can further exacerbate service delays. To address this, the introduction of an enhanced Dijkstra algorithm is proposed, aiming to identify the shortest and most efficient path between service requests and the necessary resources. It's important to note, however, that the shortest path does not automatically guarantee minimal service delay, as real-time network conditions—such as

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00065

bandwidth and resource occupancy—can also impact service delivery times. The enhanced Dijkstra algorithm, therefore, incorporates considerations for these dynamic conditions to ensure that services are delivered to users with minimal delays, thereby maintaining or improving the quality of service in distributed edge computing environments[7].

3 RL-BASED RESOURCE CACHING

In the context of resource scheduling within cooperative edge networks, the prevailing assumption often leans towards the notion that resources are adequately pre-cached at edge nodes (ENs). This presumption, while simplifying the initial stages of resource management, falls significantly short of addressing the nuanced complexities and dynamic nature of modern edge computing environments. These environments are characterized by rapidly evolving user demands and the inherent unpredictability of network conditions, which collectively exacerbate the challenges of resource allocation and optimization [1]. In light of these challenges, the imperative to design and implement more efficient resource caching methodologies becomes clear. Such methodologies are crucial for enhancing resource utilization rates, ensuring that available resources are allocated in a manner that maximally benefits the network's performance and user satisfaction [8, 10]. To tackle this challenge, we model the problem, revealing it as a mathematical programming dilemma with equilibrium constraints that exhibits non-convexity. Greedy methods frequently fail to effectively resolve these problems, often yielding only local solutions.

The proposed solution employs a deep reinforcement learning (DRL) strategy. Initially, we design a deep neural network to serve as the decision-making model for ENs. The model's input comprises the current network environment's state information and the initial caching queue, with the output being the caching queue after adjustments based on cooperative game theory. The action space determines alterations to the initial cache, based on the edge's pricing strategy and the environmental state. For training this model, a reinforcement learning algorithm such as the Deep Q-Network (DQN) is recommended. Edge profits and acquisition costs serve as the reward function in the model, evaluating each EN's decisions[14]. Upon training completion, ENs utilize this model to make decisions based on the current network status at each time slot, and store the selected caching resources accordingly. Users choose the ENs from which to receive services, guided by their requirements. Through ongoing interaction with the actual environment, the model is able to further refine its decision-making strategies, thereby enhancing overall performance. This approach aims to dynamically and intelligently manage edge caching, addressing the complex and evolving needs of cooperative edge networks and user demands, thus promising significant enhancements in resource utilization and service efficiency.

4 CONCLUSION AND FUTUREWORK

In this study, we offer a comprehensive summary of our recent and ongoing research [5, 6, 9], focusing on evaluating various strategies for service and request scheduling in AIoT. Our findings suggest that the proposed model significantly outperforms competing models. Furthermore, we have investigated various techniques within cooperative edge computing. Our findings elucidate the mechanisms and dynamics of cooperative edge computing, achieving efficient service placement and request scheduling, thereby offering computational support to AI-enabled IoT applications. In future work, we aim to develop a robust model capable of managing scenarios in which AIoT information is lost during resource caching and scheduling. Additionally, we plan to establish a testbed to demonstrate our schemes in real-world AIoT systems.

5 BIOGRAPHY

Yuzhu Liang is currently a Ph.D. candidate at the School of Artificial Intelligence, Beijing Normal University, China. He commenced his Ph.D. studies in September 2021 and is anticipated to graduate in January 2025. His doctoral research is supervised by Prof. Tian Wang, who leads the Beijing Normal University Edge AI group. His research interests encompass Edge computing and the Internet of Things.

REFERENCES

[1] Xianzhong Ding and Wan Du. 2022. Smart irrigation control using deep reinforcement learning. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 539–540.

[2] Vajiheh Farhadi, Fidan Mehmeti, Ting He, Thomas F La Porta, Hana Khamfroush, Shiqiang Wang, Kevin S Chan, and Konstantinos Poularakis. 2021. Service placement and request scheduling for data-intensive applications in edge clouds. *IEEE/ACM Transactions on Networking* 29, 2 (2021), 779–792.

[3] Bin Li, Yufeng Liu, Ling Tan, Heng Pan, and Yan Zhang. 2022. Digital twin assisted task offloading for aerial edge computing and networks. *IEEE Transactions on Vehicular Technology* 71, 10 (2022), 10863–10877.

[4] Kexin Li, Xingwei Wang, Qiang He, Mingzhou Yang, Min Huang, and Schahram Dustdar. 2023. Task computation offloading for multi-access edge computing via attention communication deep reinforcement learning. *IEEE Transactions on Services Computing* (2023). 10.1109/TSC.2022.3225473.

[5] Yuzhu Liang, Wenhua Wang, Xi Zheng, Qin Liu, Liang Wang, and Tian Wang. 2023. Collaborative Edge Service Placement for Maximizing QoS with Distributed Data Cleaning. In *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*. IEEE, 1–4.

[6] Yuzhu Liang, Mujun Yin, Yilin Zhang, Wenhua Wang, Weijia Jia, and Tian Wang. 2023. Grouping reduces energy cost in directionally rechargeable wireless vehicular and sensor networks. *IEEE Transactions on Vehicular Technology* 72, 8 (2023), 10840–10851.

[7] Junkun Peng, Qing Li, Xun Tang, Dan Zhao, Chuang Hu, and Yong Jiang. 2023. A Cooperative Caching System in Heterogeneous Edge Networks. *IEEE Transactions on Mobile Computing* (2023). doi:10.1109/TMC.2023.3336955.

[8] Tian Wang, Yuzhu Liang, Weijia Jia, Muhammad Arif, Anfeng Liu, and Mande Xie. 2019. Coupling resource management based on fog computing in smart city systems. *Journal of Network and Computer Applications* 135 (2019), 11–19.

[9] Tian Wang, Yuzhu Liang, Xuewei Shen, Xi Zheng, Adnan Mahmood, and Quan Z Sheng. 2023. Edge computing and sensor-cloud: Overview, solutions, and directions. *Comput. Surveys* 55, 13s (2023), 1–37.

[10] Tian Wang, Yuzhu Liang, Yujie Tian, Md Zakirul Alam Bhuiyan, Anfeng Liu, and A Taufiq Asyhari. 2021. Solving coupling security problem for sustainable sensor-cloud systems based on fog computing. *IEEE Transactions on Sustainable Computing* 6, 1 (2021), 43–53.

[11] Tian Wang, Yuzhu Liang, Yi Yang, Guangquan Xu, Hao Peng, Anfeng Liu, and Weijia Jia. 2020. An intelligent edge-computing-based method to counter coupling problems in cyber-physical systems. *IEEE Network* 34, 3 (2020), 16–22.

[12] Tian Wang, Yuzhu Liang, Yilin Zhang, Xi Zheng, Muhammad Arif, Jin Wang, and Qun Jin. 2020. An intelligent dynamic offloading from cloud to edge for smart iot systems with big data. *IEEE Transactions on Network Science and Engineering* 7, 4 (2020), 2598–2607.

[13] Feng Xiang, Shun Zhou, Ying Zuo, and Fei Tao. 2022. Digital twin driven end-face defect control method for hot-rolled coil with cloud-edge collaboration. *IEEE Transactions on Industrial Informatics* 19, 2 (2022), 1674–1682.

[14] Yikai Zhao, Wenrui Liu, Fenghao Dong, Tong Yang, Yuanpeng Li, Kaicheng Yang, Zirui Liu, Zhengyi Jia, and Yongqiang Yang. 2023. P4LRU: Towards An LRU Cache Entirely in Programmable Data Plane. In *Proceedings of the ACM SIGCOMM 2023 Conference*. 967–980.