Performance Analysis of Resource Allocation in 5G & Beyond 5G using AI

Niyati R. Karani[1], Jugalkumar R. Lad[2], Kinjal M. Vaghasiya[3], Pradnya V. Kamble[4]

[123]Students, Department of Electronics and Telecommunication Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India .
[4]Assistant Professor, Department of Electronics and Telecommunication Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India.

---***---

Abstract - ***In the foreseeable future, the increase in the number of devices and the Internet of Things (IoT) can make it troublesome for the latest cellular networks to make sure adequate network resources are allocated. Future network technologies have attracted increasing attention, by delivering new style ideas with dynamic resource allocation. Device resource requirements are typically variable, so a dynamic resource allocation methodology is adapted to ensure the efficient execution of any task. In this project, Performance Analysis of Resource Allocation in Future Cellular Networks with AI an attempt has been made to replicate a modern cellular data structure that supports dynamic resource allocation. Therefore, we will be designing a modern cellular data structure that supports dynamic resource allocation. Then, a dynamic nested neural network is built, which adjusts the nested learning model structure online to meet the training necessities of dynamic resource allocation. An AI-driven dynamic resource allocation algorithm (ADRA) is used that supports the nested neural network combined with the Markov decision process for training a Modern cellular data structure. The results will thus validate that the algorithm improves the average resource hit rate and reduces the average delay time.***

***Key Words*: Artificial intelligence, Edge Computing, Resource Allocation, 5G, Reinforcement Learning, Neural Network, Beyond 5G.**

1. INTRODUCTION

Cellular communication structures have progressed from rudimentary designs to cutting-edge machines during the last many decades. For the past few decades, the wi-fi communications industry has been one of the few that has maintained a fast-growing trend with unique features.

Mobile communication systems have progressed from their beginnings to the fifth generation. Increased user demand and advancements in IP technology are driving the evolution of cellular structures from first to fourth technology. Only simple and basic voice services are provided by the first-generation analog system. With the advancement of digital technology, the widely used GSM system is able to provide significantly increased capacity and facilitate international roaming. The emergence of smartphones has resulted in a significant change in the 3G network structure, both in terms of circuit switching and packet switching. The 3G system changed the way we connected dramatically. It has faster data speeds than 2G and can handle a wide range of services, including online surfing, e-mail, video conferencing, and navigation maps. However, there are three major standards in use around the world, and interoperability between them is difficult to achieve.

The fourth generation of cellular wireless technologies, known as 4G, has replaced the previous generation of broadband mobile communications. The 4G service offering's premise is to provide a comprehensive IP-based solution that allows multimedia applications and services to be supplied to users at any time and from any location, with a high data rate, the superior quality of service, and high security. The functionality of 4G communications depends on seamless mobility and interoperability with existing wireless standards.

LTE is the fourth generation of mobile communication technology, commonly known as "4G." The "Universal Mobile Telecommunications System" (UMTS) - the third generation of mobile communications – has been enhanced with LTE. Accordingly, 5G is made up of a series of improvements to LTE. As a result, 5G clearly represents the next step in the evolution of cellular technology. An IP network architecture is at the heart of the LTE network. As a result, Deutsche Telekom's fixed-line network has also been converted to Internet Protocol (IP).

5G will not only support communication, but also processing, management, and content delivery (4C) functions, as compared to earlier generations. With the introduction of 5G, several new applications and use cases are predicted, such as virtual/augmented reality (VR/AR), unmanned cars, Interactive Net, and IoT applications. These applications are expected to result in significant growth for both communication and computational resources.

Wireless network systems rely heavily on resource allocation. To meet different network requirements in a 5G communication network, the system must be smarter and more dynamic in nature. The system allocates resources for power control, bandwidth allocation, deployment techniques, and association allocation. In any cellular network architecture, resource allocation is critical. It is crucial in ensuring that end-users, business partners, and customers of cellular-based applications have easy access. The cellular network environment benefits greatly from resource allocation. The level of fairness of the network's resource allocation determines network performance. The level of fairness has a strong link to the network's performance. The resource allocation fairness levels are fair, ideal, unfair, and unbalanced. The performance levels of the network are poor, less good, good, and perfect.

Our main aim is to design a dynamic resource allocation model for the latest cellular network where the density of users is bigger than the density of available resources by considering both the average resource hit rate and the average delay time of the network.

On the other hand, running distributed machine learning model training at the edge of the network is valuable for solving resource allocation problems under different optimization goals. We established a resource allocation neural network model for a modern cellular structure to achieve the high real-time performance of dynamic resource allocation, which is also in line with the inevitable trend of the development of large-scale IoT applications. Finally, an AI-driven dynamic resource allocation algorithm is designed to realize appropriate resource allocation optimization in the latest cellular networks.

2. RELATED WORKS

Research carefully associated with this text is in particular divided into the subsequent 3 categories:

1) Future Cellular Network

2) Intelligent Edge Computing 3) Resource Allocation.

2.1 FUTURE CELLULAR NETWORK

The communication industry is anticipated to grow at an incredible rate, encouraging innovation and productivity in all other economic sectors like transportation, health care, agriculture, finance, services, consumer electronics, and so on.[11] According to Zhang et. al., the Internet of Things (IoT), vehicle-to-everything (V2X), distance learning, (advanced) virtual and augmented reality, unmanned aerial vehicles (UAVs), and robotization are all new concepts that can significantly impact our lifestyles in the coming years.[11] So the need for AI in 5G/B5G development needs domain-specific knowledge in communication technology, proficiency in Machine learning and artificial intelligence, and hardware design experience.[11]

2.2 INTELLIGENT EDGE COMPUTING

Traditional cloud computing is not the same as edge computing. It's a modern computing paradigm in which computation takes place at the network's edge.[5] Its primary concept is to bring computational processing closer to the data source.[5] Edge computing is defined differently by different researchers. Shi et al. [12] introduced the emergence of the concept of edge computing: ''Edge computing is a new computing mode of network edge execution. The downlink data of edge computing represents cloud service, the uplink data represents the Internet of Everything, and the edge of edge computing refers to the arbitrary computing and network resources between the data source and the path of cloud computing center.''[12] Edge computing, in other words, is the provision of services and computations at the network's and data generation's edge.[12] Edge computing refers to the migration of the cloud's network, computing, storage, and resource capabilities to the network's edge, as well as the supply of intelligent services at the edge to satisfy the necessities of the IT industry in agile linking, real-time business, data optimization, application intelligence, security, and privacy, as well as the network's low latency and high bandwidth requirements.

2.3 RESOURCE ALLOCATION

To make complete use of side and terminal gadgets, deep gaining knowledge of fashions tends to be allotted on a couple of gadgets.[9] However, adjustments in undertaking necessities and variations in gadgets have introduced top-notch challenges, mainly to a pressing want for cheap aid allocation strategies as a prerequisite for making sure undertaking performance.[13] A large number of edge devices, such as sensors and actuators, has led to an enormous increase in data processing, storage, and communication requirements. A cloud platform has been proposed for connecting a massive number of IoT devices, with a huge amount of data generated by such devices being offloaded to a cloud server for processing. Edge devices, as opposed to cloud servers, can provide latency-critical services and a variety of IoT applications. End devices are resource-constrained in general; for instance, battery capacity and internal CPU computation capacity are both limited [13]. Offloading computational activities to relatively resource-rich edge devices can satisfy application quality of service (QoS) needs while also improving end-device capabilities for resource-intensive applications.[13]

3. SYSTEM ARCHITECTURE

3.1 PROBLEM FORMULATION

In 5G, high throughput and ultra-low communication latency are proposed to be achieved, to improve users' quality of experience (QoE). For this, 5G targets three evolution axes to cope with the new applications fields; such as autonomous cars/driving, industrial automation, virtual reality, e-health, etc. To achieve this in a highly dense populated area where the available resources are comparatively lower than the need of required resources by the users.

Thus, in a massive IoT architecture, it is necessary to implement a machine learning algorithm for better performance for the task of resource allocation. Since the task request and user cases may vary with respect to time we need to implement a real-time-based simulation environment and a real-time-based AI algorithm for better performance. Thus, we have implemented a dynamic neural network framework in this project.

3.2 IMPLEMENTATION

A dynamic resource allocation model for the latest cellular network where the density of users is bigger than the density of available resources is designed and then its performance is analyzed in terms of hit rate and delay time. In this project, we have divided work into two blocks namely, the resource selection block, and the neural network block.

A. RESOURCE SELECTION STAGE

In the first block, i.e. resource selection block is subdivided into three layers 1) the processing layer; 2) the object layer; and 3) the distance layer. In the processing layer, the process or a given task is simulated. Here, we have used a single task to simplify the process of simulation. This task type is of "Objection detection: VOC SSD300" with a processing size of 300 * 300 * 3 * 1 bytes, loading size of 300 * 300 * 3 * 4 bytes and transmission data size of 4 * 4 + 20 * 4 bytes.

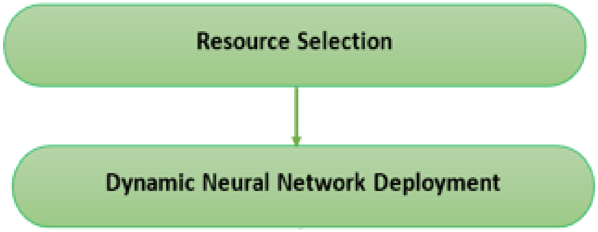


Figure 1- Dynamic allocation process

Then, in the object layer, we simulate the user data, the user can be mobile or a stationary device. Here, a massive IoT architecture is simulated i.e. the Edge Server is responsible for offering computational resources (5.04 * 108 bits/sec) and the process request states are generated for multiple use-cases. For simplification of the project, six states are defined namely state one - offloading of a task to the edge server, state two - sending of task request to the edge server (2.16 * 105 bits), state three - processing of the requested task (8.64 * 106 bits), state four - the processed task results are sent back to the user device (768 bits are required), state five- disconnect state (default), state six- migration of task to another edge server.

In the distance layer, the mobility data of the user is simulated using the data i.e KAIST Data set provided by CRAWDAD. The mobile devices of a subway station in Korea from the city of Seoul are collected to form the KAIST data set.

B. DYNAMIC NEURAL NETWORK DEPLOYMENT STAGE

Let's first understand the problem in allocating computer resources and migration bandwidth, while selecting the offloading server for each user is a discrete variable challenge. As a result, for continuous and discrete problems a non-policy, non-model, and actor-critic network structure-based Deep Deterministic Policy Gradient (DDPG) method is used. DDPG updates the model weights at every step, thus, allowing the model to immediately adapt to changes in the environment.

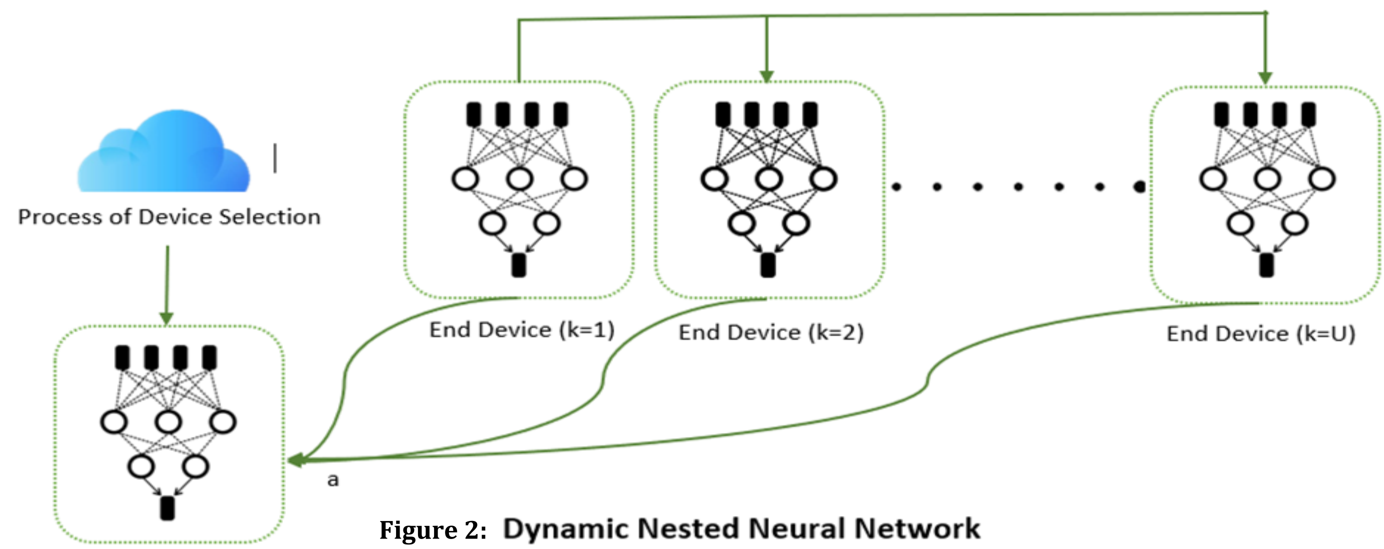


Figure 2: Dynamic Nested Neural Network

Here, in this structure, two dynamic neural networks are implemented.

- Actor - It proposes an action given a state.
- Critic - It predicts if the action is good (positive value) or bad (negative value) given a state and an action.

In the actor-network, a four-layer neural network is implemented for each user. Here, user states are simulated and their required resources are calculated. Then, using MDP the results of this network are forwarded to the Edge Server and on these required resources another two four-layer neural networks are trained for the processing layer - first for the resources selection and second for the use of bandwidth. Here, the resources are deployed to the user as per their requirements, and its efficiency is tried to increase by using neural networks of the simulated action space.

In the critic network, the action space and its efficiency are used as input to form a backtracking network structure, thus helping in predicting whether the action is a good action or a bad action for a given state.

This, critic network output is in turn used in the action network as input for the next training episode thus, successfully helping the real-time model to update and train itself for better efficiency.

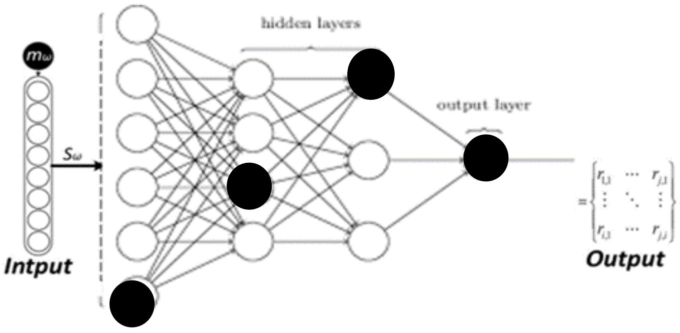


Figure 3- Resource allocation training process of Algorithms.

4. PERFORMANCE ANALYSIS AND RESULTS

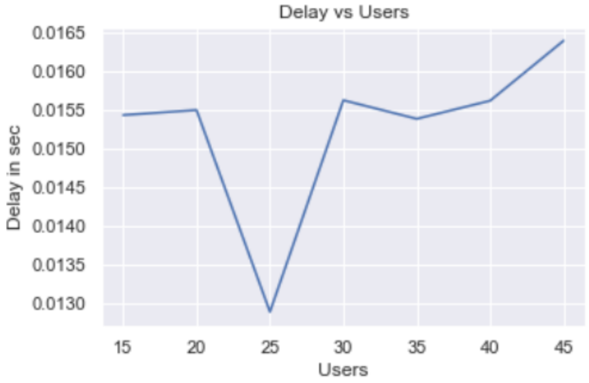
4.1 PERFORMANCE ANALYSIS

In this section, the proposed algorithm is evaluated for various users namely (15, 20, 25, 30, 35, 40, and 45 users) and 3 edges in the distance of 3.5 km having a resource limit of up to 5.04 * 108 bits/sec. In addition, the user neural network is randomly deployed on each device. The key objective of the simulation is to see if the proposed algorithm can make real-time resource allocation decisions to ensure that task-executing devices have the resources when they need to complete the task smoothly. Since it is performed under 5G the communication rate is set to 5GHz/s and the delay is set to 0.1 ms. The hit rate is defined as the measure of the accuracy of the final decision result, i.e it is the ratio of the resource actually allocated to the device and the resource demand. The delay time is defined as the time required to successfully execute the given task i.e. it is the measure of the difference between the time at which the task process is requested by the device and the task process is completed and is transmitted to the device.

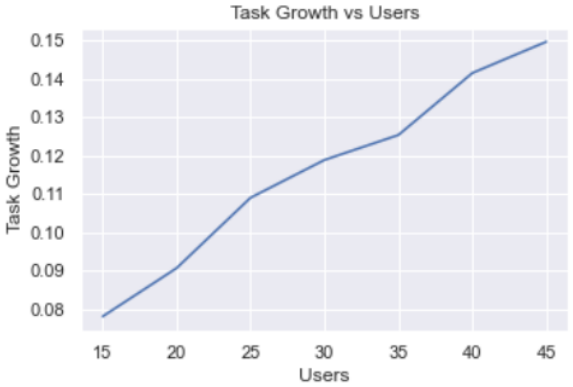
For the nested neural network, its number of dimensions is the same as the number of devices selected to nest. Therefore, only the devices with local neural networks are selected. The threshold of these local neural networks is an important factor that determines the performance of the nested neural network,

4.2 RESULTS OBTAINED

The figure shows the average decision delay time, which is the average waiting time of the device from initiating the request to obtaining the decision result. It should be noted that this online decision delay is based on previous training results instead of a complete training process from scratch. The comparison is done by changing the device distribution density. The devices with resource requests are randomly selected. It can be found that with the increased scale of device distribution density, although the training time of all algorithms increases, the decision time of the algorithm is more or less in the range of 0.013 to 0.017 secs.



The figure shows the average task growth percentage with respect to the number of users, it can be seen that as we increase the number of users the percentage of tasks completed increases. It should be noted that it is not the difference in the average task before and after training but its percentage.Thus it can be scaled on the same graph.



The average resource hit rates are shown in Figure, where the time period is set as 0.25 min to observe the real-time performance among them. We can see that at each point in time the average resource hit rate seems to be gradually increasing.

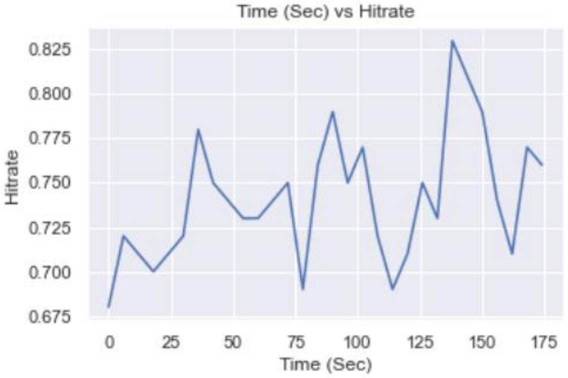


Table 1 - Average Decision delay time, hit rate and task completed under different devices

	Tasks Completed	Hit Rate	Delay Time
15 users	15450.534	0.6504	0.015
20 users	20438.867	0.665	0.015
25 users	25611.2	0.741	0.013
30 users	30550.7	0.594	0.016
35 users	35289.9	0.745	0.015
40 users	40546.934	0.626	0.016
45 users	45439.0	0.584	0.016

5. CONCLUSIONS AND FUTURE SCOPE

In this model, we haven't taken into consideration the uncertainty in the channel state information. This work can be extended by incorporating the uncertainty in the wireless propagation environment.

Our project only focuses on maximizing the total network throughput. This work can be extended by incorporating fairness or users' QoS requirements (e.g. minimum rate guarantee for each user) in the resource allocation model. Unfortunately, these networks will be interference-limited, since orthogonal transmission schemes and/or linear interference neutralization techniques are not practical due to the massive amount of nodes to be served.

We can consider other parameters, such as energy consumption and mobility, to improve the mechanism.

6. REFERENCES

[1] H. Ye, G. Ye Li, and B.-H. F. Juang, "Deep reinforcement learning-based resource allocation for V2V communications," IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[2] "K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI-empowered the wireless networks," IEEE Commun. Mag., vol. 57, no. 8, pp. 84–90, Aug. 2019.

[3] M. Chen, Y. Miao, H. Gharavi, L. Hu, and I. Humar, "Intelligent traffic adaptive resource allocation for edge computing-based 5G networks," IEEE Trans. Cognitive Commun. Netw., vol. 6, no. 2, pp. 499–508, Jun. 2020.

[4] S. Wang, T. Sun, H. Yang, X. Duan, and L. Lu, "6G network: Towards a distributed and autonomous system," 2020, pp. 1–5.

5] Zhiyong Liu∗, Xin Chen∗, Ying Chen∗, Zhuo Li, "Deep Reinforcement Learning Based Dynamic Resource Allocation in 5G Ultra-Dense Networks", 2019 pp.

[6] Kai Lin , Senior Member, IEEE, Yihui Li, Qiang Zhang , and Giancarlo Fortino , Senior Member, IEEE, "AI-Driven Collaborative Resource Allocation for Task Execution in 6G-Enabled Massive IoT",2021 pp.

[7] Y. Dai, K. Zhang, S. Maharjan, and Y. Zhang, "Edge intelligence for energy-efficient computation offloading and resource allocation in 5G beyond," IEEE Trans. Veh. Technol., vol. 69, no. 10, pp. 12175–12186, Oct. 2020.

[8] K. Lin, C. Li, Y. Li, C. Savaglio, and G. Fortino, "Distributed learning for vehicle routing decision in software defined Internet of vehicles," IEEE Trans. Intell. Transp. Syst., early access, Sep. 28, 2020

[9] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentization mobile edge computing, caching and communication by federated learning," IEEE Netw., vol. 33, no. 5, pp. 156–165, Sep./Oct. 2019.

[10] Chen, Z. Li, K. Wang, and L. Xing, "MDP-Based Network Selection with Reward Optimization in HetNets," Chinese Journal of Electronics, Vol.27, No.1, pp. 183-190, Jan. 2018.

[11] C. Zhang, Y. -L. Ueng, C. Studer and A. Burg, "Artificial Intelligence for 5G and Beyond 5G: Implementations, Algorithms, and Optimizations," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 10, no. 2, pp. 149-163, June 2020, doi: 10.1109/JETCAS.2020.3000103.

[12] K. Cao, Y. Liu, G. Meng and Q. Sun, "An Overview on Edge Computing Research," in IEEE Access, vol. 8, pp. 85714-85728,2020, doi: 10.1109/ACCESS.2020.2991734