Beyond 5G White Paper
Supplementary Volume
“AI/ML Technologies”

Version 1.0

March 7, 2024

Beyond 5G Promotion Consortium
White Paper Subcommittee

【Revision History】

Ver.	Date	Contents	Note
1.0	2024.3.7	Initial version	

Preface

The technological evolution of communication networks is rapidly progressing towards the Beyond 5G era, and artificial intelligence (AI) and machine learning (ML) technologies will play a significant role in this evolution. These technologies will be utilized in various areas to enhance the capabilities of Beyond 5G. Specifically, AI/ML will be used for automation of network operations and management, analysis of the characteristics of communications at each layer, optimization of computing and networking resources, and adaptation to meet the individual requirements of diverse services/applications. When utilizing these AI/ML technologies, in addition to replacing conventional technologies with learning-based technologies, it is necessary to develop AI/ML technologies that incorporate the characteristics and knowledge of physical and mathematical models that have been developed over the years.

This white paper introduces leading-edge R&D efforts on AI/ML for Beyond 5G in Japan, with the aim of accelerating R&D to advance future communications and services. This supplementary volume on AI/ML technologies includes six papers categorized into three typical categories, as shown below:

AI/ML for Network Operations

1. AIOps for Autonomous Networks
2. Logic-Oriented Generative AI Technology for Autonomous Networks

AI/ML for Radio Access Management

3. Scalable AI/ML for Radio Cellular Access
4. AI/ML-based Radio Propagation Prediction Technology

AI/ML for User/Application-Centric Communications

5. 6G Network AI Architecture for Everyone-Centric Customized Services
6. AI-based Application-aware RAN Optimization

This white paper was prepared with the generous support of many people who participated in the White Paper Subcommittee. The cooperation of telecommunications industry players and academia experts, as well as representatives of various industries other than the communications industry has also been substantial. Thanks to everyone's participation and support, this white paper was able to cover a lot of useful information for future business creation discussions between the industry, academia, and government, and for investigating solutions to social issues, not only in the telecommunications industry, but also across all industries. We hope that this white paper will help Japan create a better future for society and promote significant global activities.

Eiji Takahashi, NEC

AIOps for Autonomous Networks

Takuya Miyasaka, KDDI Research, Inc.
Minato Sakuraba, KDDI Research, Inc.
Tananun Orawiwattanakul, KDDI Research, Inc.
Atsushi Tagami, KDDI Research, Inc.

Abstract— **This report provides an overview of Autonomous Networks expected to be realized in Beyond 5G. Furthermore, this report describes the details of network operation by AI, which is a necessary element of the Autonomous Network, and especially summarizes the strategy for managing network failures, and provides the overall framework required for future network operation.**

1. Introduction

The fundamental role of the mobile network is to provide connectivity for user equipment (UE). Furthermore, to achieve high-quality mobile service, the network must meet the quality requirements of UE and the web services with which UE communicates. In traditional network operations, human operators have played this role. Operators install the network equipment, such as base stations and servers, that constitutes the mobile network, configure them appropriately, and replace them in the event of failures. These critical tasks enable the mobile network to meet these quality requirements around the clock.

In recent years, there has been a lot of standardization, research, and development activities on Autonomous Networks [1,2], where the network autonomously performs these tasks traditionally performed by human operators. As shown in Fig. 1, in the Autonomous Network, the network's configuration and control are managed autonomously based on Intent information, which represents the requirements of actual users of the network.

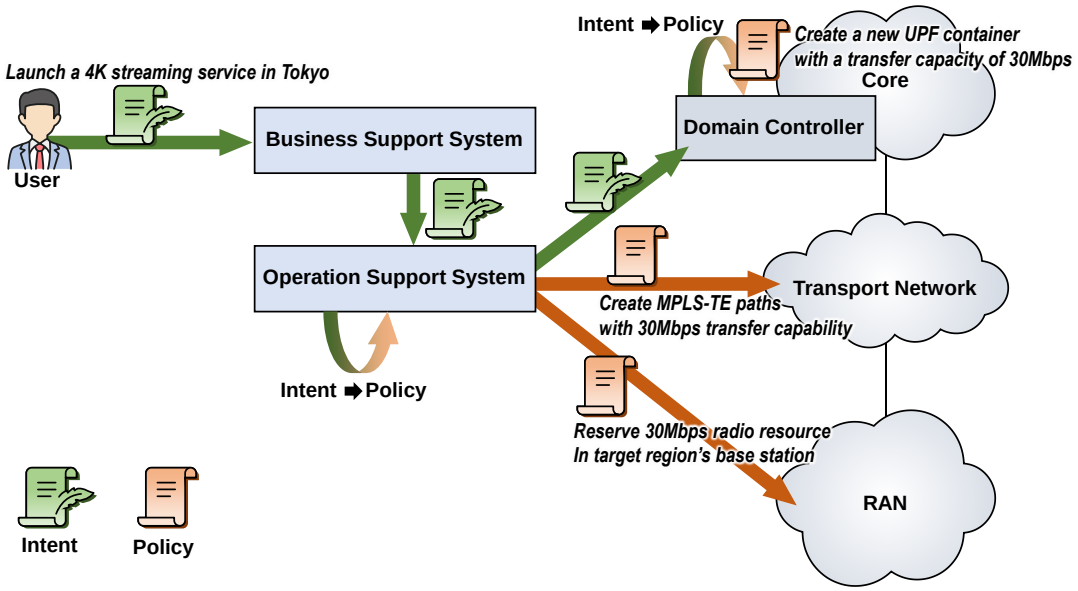


Fig. 1 General concept of Autonomous Network

Intent is more abstract information than policy, rules, and logic regarding the network and represents an intention and expectation of the network's user. In the example in Fig 1, the operational system, which consists of the Business Support System (BSS) and the Operation Support System (OSS), receives an Intent from a user who wants to launch a 4K streaming service in Tokyo, divides the Intent into each network domain, translates it into an actual network control policy, and requests it to each network domain. In some cases, Intent may also be sent directly to a domain controller that controls each network domain without being translated into a policy in the operational system. Since the domain controller has a more detailed understanding of the operational data of each network domain, a more detailed and accurate policy translation can be expected. Each domain controller implements control of the relevant network to ensure the quality specified in the policy.

2. AIOps and Autonomous Networks

Artificial Intelligence for IT Operations (AIOps) is essential for achieving Autonomous Networks. As described in the above section, in Autonomous Networks, it is necessary to translate abstract Intent received from users, e.g., "*I want to launch a 4K streaming service in Tokyo*", into concrete policies and rules, e.g., "*Creating MPLS-TE paths with 30~Mbps transfer capability*". Intent allows different users to request network services without using a technical language that they do not usually use, such as a programming language. However, user Intent varies widely, making traditional fixed rule-based translation difficult. Furthermore, it is essential to build the network

policy translated from the Intent on the network infrastructure (RAN, Transport Network, and Core) and to deal with network failures without human operators. To address such issues, AIOps for Autonomous Networks requires three key elements: 1. Intent translation, 2. Network resource management, and 3. Network failure management.

In 1. Intent translation, users' abstract Intent is translated into a specific network policy. Generative AI and the Large Language Model (LLM), which have been actively researched and developed for practical use in recent years, can be applied to this process. Furthermore, the interaction between a user and AI is beneficial not only for understanding the user's needs but also for negotiating with the user, for example, negotiating alternative proposal by AI when network resources are insufficient.

In 2. Network resource management, based on the converted network policy, network resources are reserved, and the user-requested network service is created and provided to the user. An optimal resource allocation placement is determined to satisfy the network policy, and network elements (e.g., virtual mobile core, MPLS-TE path, virtual CU/DU) that constitute the user's network service are generated on demand. In addition, network resources are not always prepared enough to always satisfy all user requests and accommodating them may not be possible. In such cases, admission control of user requests is necessary, and based on the request status (new requests, cancellations), decisions must be made to maximize the profit of the network operator, and automated decision-making, such as Deep Reinforcement Learning, can be applied [3].

Finally, in 3. Network failure management, when a network failure (e.g., HDD failure, link down, restart) occurs in the created user network service, a series of processes that detect the failure event, identify the root cause, and resolve the issue are implemented. Various AI technologies, such as anomaly detection and classification, are being considered and introduced. The next section of this report describes the detailed technical aspects obtained through our research results.

3. Network Failure Management by AIOps

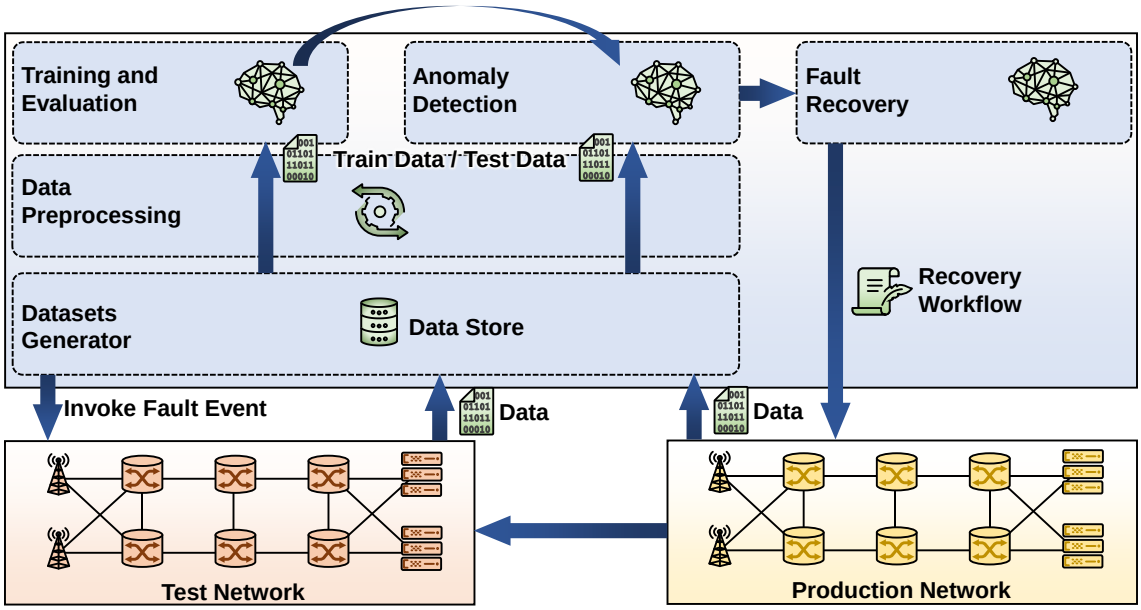


Fig. 2 AIOps framework for Network Failure Management

Our proposed integrated framework [4] for network failure management with AIOps is shown in Fig. 2, including data collection, anomaly detection, and fault recovery functions. The framework has three phases: the data collection for AI model training, the AI model training phase for AIOps, and the AI model inference phase from actual operational data. Firstly, in the data collection phase, the network devices, such as servers and routers, that consist of the operation target network send statistical data. Typical statistical data include CPU, memory, network, and other resource utilization rates. Furthermore, user utilization data (e.g., # of sessions) related to mobile network software such as PGW and UPF is also included. In addition, we have proposed a method for anomaly detection and prediction based on Observability with Linux eBPF, which is frequently used in cloud-native environments [5,6]. Since data is essential for training highly accurate AI models, more detailed data describing system behavior, such as eBPF, will be required in future mobile networks.

Secondly, in the AI model training phase, AI models are trained from operational data, such as CPU utilization rate and trouble tickets, obtained in the target network. Since network failures are infrequent events in production networks, sufficient operational data for AI models may not be obtained. Therefore, a test network simulating the production network can be created to train precise AI models, and operational data obtained from pseudo network failure generated in the test network can be utilized as input data for the AI model.

Finally, in the AI model inference phase, the trained AI model detects a root cause of network failure from the latest operational dataset and suggests an optimal recovery workflow from the network failure, with anomaly detection and fault recovery function. The anomaly detection function detects network failures and determines their root causes. Within this framework, we have evaluated a comparative experiment that involved measuring the performance of the fault analysis function using three AI algorithms, multi-layer perceptron (MLP), random forest (RF), and support vector machine (SVM), on the testbed network built by the virtualized network functions (VNFs) [7]. RF showed the highest accuracy, and F1 scores for three network failures: compute node down, network interface down, and CPU overload were 1.00, 0.96, and 0.95, respectively. This difference in accuracy by AI algorithms is likely due to the dataset generated from the performance management (PM) data, and the increase in training data, feature reduction, or balance adjustment of normal/abnormal samples affected the accuracy.

Furthermore, we have proposed a scheme for fault recovery using reinforcement learning (RL) [8]. The scheme can adapt to network topology and configuration changes and has a data representation procedure to prepare a data set for RL, which is formed as a matrix of network topology and fault state. The simulation results showed that preparing enough training data requires a tremendous amount of failure injection and recovery operation trials. The test network simulating the production network can potentially shorten the time for trials in the training process. However, our simulation also revealed that the behavior between the test network and the production network infrastructures should be 87% coincident for application to the proposed scheme.

4. Conclusion

This report described an overview of Autonomous Networks and AIOps. To benefit from the convenience brought by Autonomous Networks, it is necessary to introduce the concept of such Autonomous Networks and AIOps as the network architecture for Beyond5G system. More specifically, it is essential to have architectural support to create an end-to-end network instance and control user policies on the network instance based on user Intent. Furthermore, the Beyond5G system also needs to centrally manage operational data from the RAN, Core, and Transport Network in an integrated way and automatically train and deploy the optimal AI model for AIOps.

REFERENCE

[1] A. Boasman-Patel, *et al.*, “Autonomous Networks: Empowering Digital Transformation for the Telecoms Industry,” White Paper, TM Forum, May 2019.
[2] 3GPP, “Technical Specification Group Services and System Aspects; Management and orchestration; Levels of autonomous network,” 3GPP TS 28.100, ver. 17.1.0, September 2022.
[3] T. Orawiwattanakul, et al. "Reinforcement Learning (RL) Based Admission Control in Advance Bandwidth Reservation." IEEE/IFIP Network Operations and Management Symposium (NOMS), 2024 (To be published).
[4] A. Tagami, et al. "Integration of Network and Artificial Intelligence toward the Beyond 5G/6G Networks." IEICE Transactions on Communications 106.12 (2023): 1267-1274.
[5] J. Kawasaki, et al. "Failure Prediction in Cloud Native 5G Core With eBPF-based Observability," 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), Florence, Italy, 2023, pp. 1-6, doi: 10.1109/VTC2023-Spring57618.2023.10200028.
[6] M. Sakuraba, et al. "An Anomaly Detection Approach by AIML in IP Networks with eBPF-Based Observability," 2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS), Sejong, Korea, Republic of, 2023, pp. 171-176.
[7] J. Kawasaki, et al. "Comparative analysis of network fault classification using machine learning." NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2020.
[8] T. Miyamoto, et al. "Network Topology-Traceable Fault Recovery Framework with Reinforcement Learning," Proc. of Advanced Information Networking and Applications, AINA, pp.393–402, April 2021.

Logic-Oriented Generative AI Technology for Autonomous Networks

Takayuki Kuroda, NEC

Abstract—**Autonomous network operation technology based on intent has been attracting attention toward advanced automation of network operation. However, the realization of intent translation, which is the key to this technology, faces the challenge of achieving both flexibility and faithfulness. In this paper, we introduce a method to realize a logic-oriented generative AI, in which a logical search engine is enhanced by AI/ML technology, in intent translation. The paper presents the position of this technique with respect to related techniques, and then briefly outlines the method.**

1. Introduction

Increasingly complex networks are becoming increasingly difficult to provide quickly and reliably by manual means, and a high degree of automation of operations is required [1][2]. Intent-based networking is a promising foundational approach to automate network operations [3]. Intent is information that expresses requirements in an abstract and declarative manner. According to intent-based automation techniques, a machine interprets the intent and performs the construction and operation of the network. This allows users to easily build the desired network by simply entering high-level requirements without having to enter detailed information.

To realize such a technology, the ability to translate intent into concrete network configurations is essential. Conventional techniques for intent translation are known to be based on deductive engines [5]. Another possible approach is to use an inductive inference function, such as LLM, which has recently emerged. However, the former has the problem of requiring manual addition of conditions and rules to increase flexibility. The latter has been pointed out to have a problem of faithfulness [6]. Therefore, we propose a mechanism that combines a deductive engine and an inductive AI so that the engine can search for effective solutions from a large solution space at high speed, thereby achieving both flexibility and faithfulness. In this paper, we describe the challenges of existing methods and outline the proposed technique.

2. Automation of Intent Translation and its Challenges

Intent-based networking is a new technology that offers an abstraction layer and emphasizes on the desired outcome of the network service, instead of prescribing how it is configured [4]. There are various issues that need to be addressed to realize this technology, including the means to appropriately express the various network-related intent, the means to disambiguate them, and the means to concretize the abstract

intent so that they can be deployed in practice. Among these issues, translation is the key to intent-based networking, which derives concrete network configurations from abstract intent. Figure 1 shows an overview of intent translation. The intent in this research consists of functional and non-functional requirements for the network and/or network function to be constructed. The intent translator is based on such intent information and complements it by concretizing the details necessary for the function to work.

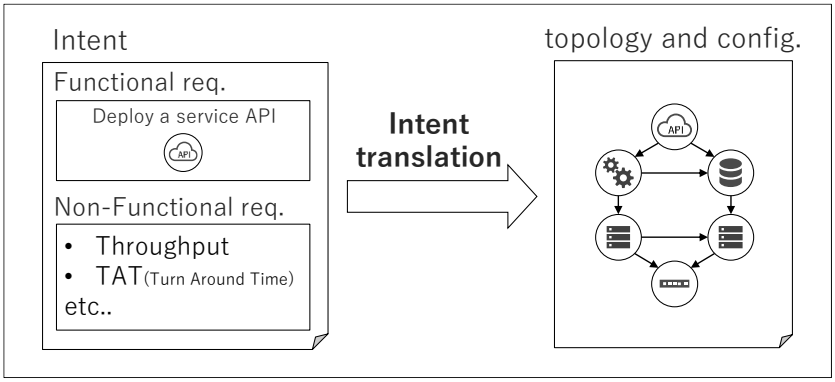


Fig. 1 Concept of automated network intent translation.

Two typical approaches to realize intent translation can be considered: deductive and inductive. In the deductive approach, the technique described in [5], the intent is refined step by step by applying predefined patterns, and a reasonable proposal of network configuration that satisfies the intent is searched among the possible proposals that can be generated. Flexibility is generally an issue with such a technique. That is, the solution is limited to specific patterns defined in advance. Although a variety of solutions can be generated by combining the patterns, it is necessary to manually align the rules to select a reasonable proposal from among them. In contrast, inductive approaches, such as the Large Language Model (LLM) that has emerged in recent years, can be utilized. Using LLM, it is expected that some answers can be obtained for any intent. However, LLMs are known to often give wrong answers and are not particularly good at thinking that involves logic, such as network design [6].

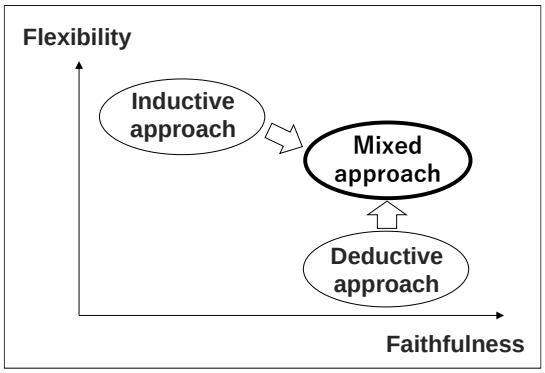


Fig. 2 Position of this study in comparison with existing studies.

Thus, a deductive approach has faithfulness but lacks flexibility, while an inductive approach has extremely high flexibility but has problems with faithfulness. In contrast, our method appropriately combines these two approaches to establish a method that achieves both sufficient flexibility and faithfulness. The relationship between each approach is shown in Fig. 2.

3. Intent Translation with Logic-oriented Generative AI

The left side of Fig. 3 shows our proposed concept of intent translation with logic-oriented generative AI. The method is based on a deductive engine, whose search is guided by a GNN-based AI, which we refer to as the design AI. The deductive engine repeatedly refines the intent in stepwise manner. At each step, the design AI evaluates the multiple proposals generated and selects the most promising proposal as the next proposal to be refined. The learning of the design AI can be performed by a reinforcement learning algorithm. An overview is shown on the right side of Fig. 3. It learns the promise of a configuration proposal by generating expected returns based on the values obtained by evaluating the results of the design trials. As the learning proceeds, we can observe an increase in the success probability of the trials. The learning process is terminated when the improvement in the learning success probability comes to a head.

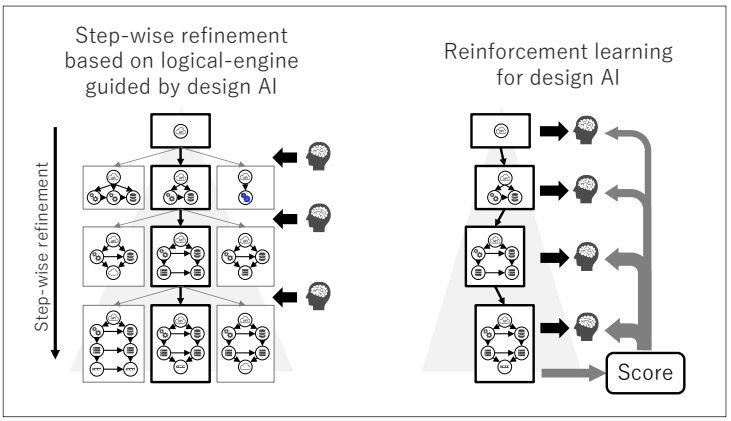


Fig. 3 concept of automated intent translation and its learning.

Design AI allows the actual search space to be narrowed down to allow flexible discovery of promising solutions from a vast potential space. In other words, the rules for searching for solutions, which were previously defined manually, are replaced by learning models.

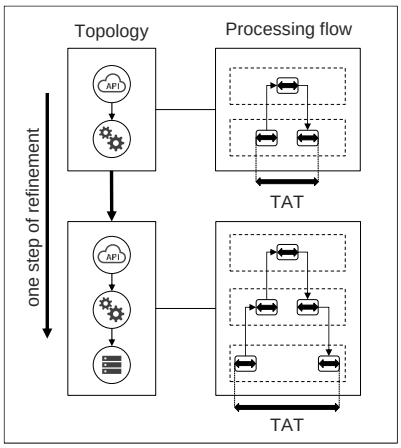


Fig. 4 Example of one step refinement of network intent and evaluation model.

The validity of the design results of this technique is guaranteed by the constraints. Figure 4 shows how, together with a network topology, the processing flow to perform the functions of the network is embodied in one step. For example, a Turn Around Time (TAT) non-functional requirement is evaluated by summing the time for each process based on the processing flow and verifying that it falls within the time specified as intent.

However, such a verification can only be performed once the design has been fully concretized. This fact has been a factor prolonging the search time, but the use of design AI can solve this problem. In the search process, it is important to make the right choice in the early stages as much as possible. This is because if a wrong decision

is made in the initial stage, many trial and errors will have to be made again. In other words, the initial decision has a larger search space for later stages. However, as Fig. 4 shows, TAT cannot be accurately determined until the last step. The incomplete processing flow shown in the upper right corner of Fig. 4 does not include some of the processes, and TAT cannot be calculated correctly. In other words, it is difficult to efficiently search a huge search space using logic alone. Instead, the design AI estimates the final TAT value from an early stage of the design process. This allows the search to be properly guided. On the other hand, constraints are essential to validate the obtained design results and to calculate accurate rewards during training.

4. Conclusion

In this paper, we introduced a technology to realize intent translation, which is a key element of intent-based networks. In particular, we described a logic-oriented generative AI that uses AI/ML technology to enhance the logical search engine in the design of network configurations to achieve both flexibility and faithfulness. In the future, we will continue to refine the technology and make it practical, as well as develop methods for accelerating learning and automating model development.

REFERENCE

[1] TM Forum. Autonomous Networks: Empowering Digital Transformation For Smart Societies and Industries. TMForum White Paper, 2020.
[2] ETSI, “Intent driven management services for mobile networks”, TS 128 312 V17.0.1 (3GPP TS 28.312 version 17.0.1 Release 17), Jul. 2022.
[3] A. Clemm, L. Ciavaglia, L. Granville, and J. Tantsura, “Intent-Based Networking Concepts and Definitions”, ITU, Geneva, Switzerland, Feb. 2021.
[4] A. Leivadeas and M. Falkner, "A Survey on Intent-Based Networking," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 625-655, Firstquarter 2023, doi: 10.1109/COMST.2022.3215919.
[5] N. Nazarzadeoghaz, F. Khendek, and M. Toeroe, “Automated design of network services from network service requirements”, in Proc. 23rd Conf. Innov. Clouds Internet Netw. Workshops (ICIN), 2020, pp. 63–70.
[6] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch, “Faithful chain-ofthought reasoning”, 2023. arXiv:2301.13379. https://doi.org/10.48550/arXiv.2301.13379

Scalable AI/ML for Radio Cellular Access

Andres Arjona, Nokia
Hideaki Takahashi, Nokia

Abstract— **Wireless networks are expected to move towards self-sustaining networks in 5G-Advanced and in 6G, where Artificial Intelligence (AI) and Machine Learning (ML) play a critical role in maintaining high performance in dynamically changing environment. AI/ML solutions that operate separately at the device or network side, or jointly on both will emerge. Similarly, lifecycle management procedures will be needed to enable interoperable automation in the radio, providing a framework with the necessary tools for deploying and operating ML solutions in radio at scale.**

1. Introduction

We are at the beginning of a revolution in cellular networks as Artificial Intelligence (AI) and Machine Learning (ML) for the air interface become integral to cellular networks. Although AI/ML is already part of 5G systems, it is currently mostly applied to network automation and proprietary Self Organizing Networks (SON) solutions. With the advent of 5G-Advanced, and further with 6G, we will see an advanced implementation of AI/ML in the RAN and radio interface. The potential benefits of AI/ML in the network will be significant. They will boost the performance of the radio interface, reduce power consumption, greatly improve the end user experience, and help find better performing network parametrization faster. Further, these solutions must be both economically and technically feasible to scale.

In this paper, we present discussion on the importance of standardizing lifecycle management procedures relevant to AI/ML, followed by an example of an AI/ML based reinforcement learning solution for uplink power control in cellular networks.

2. Lifecycle Management for AI/ML

AI/ML solutions for the air interface [3] can be one-sided, where a given feature operates at either the network or device side (e.g., beam prediction, positioning), or two-sided, where the solution operates jointly in both simultaneously (e.g., device channel feedback compression). In this latter example, the ML algorithm is applied at both the device and network side for compression and decompression of the channel state information.

Fig. 1 One-sided and two-sided AI/ML solutions in the mobile network air interface [3]

Standardization efforts are essential to ensure that different vendors' ML implementations and algorithms for networks and devices can work together in a variety of scenarios. Thus, a holistic framework needs to be developed in 3GPP for 5G-Advanced, addressing both kinds of AI/ML solutions (one-sided and two-sided) supporting control-plane signaling between the network and the device for correct and controllable operation. This framework shall be applicable to any use case in the air interface, and also be the foundation for the AI-native air interface in 6G.

Specifically, Lifecycle Management (LCM) procedures to enable interoperable automation mechanisms in the radio are needed. Including procedures for data collection, development and testing, deployment, and operation and monitoring of ML solutions. This framework will provide operators, devices, and network vendors the required tools for operating ML solutions for radio at scale with guaranteed interoperability.

Data needs to be collected for training, inference and performance monitoring of the ML solutions. Hence, the framework must ensure that operators have control about how, what, when, and for which use cases data is collected responsibly, and in compliance with local data and privacy regulations. However, a challenge for the ML training data, is regarding scalability and access to the data needed in a controlled and efficient manner. To this end, the following principles should be followed for training data collection procedures:

- Ensure user security and privacy
- Make data accessible by the subscribed parties
- Operator needs to be aware of and control data collection
- Minimize additional air-interface traffic
- Design for extensibility and future evolution

3. Deep Reinforcement Learning for Uplink Power Control

One important trend in ongoing 6G research is the paradigm shift toward self-sustainable networks. To this purpose AI and ML technologies can become key components in maintaining network performance.

Reinforcement Learning (RL) is one field in machine learning for decision making that can be applied to cellular networks. Use of RL methods can enable use cases in wireless communications and radio resource management which are otherwise difficult due to the complex nature of the radio environment. In RL, the objective is to have an agent have freedom to learn a solution, where learning of the decisions is carried out via an arbitrary function that maximizes a "reward". Throughout this process the agent learns from the reward feedback signal, which reinforces the desired actions and penalizes the undesired ones. The agent interacts with the environment by taking an action based on the observed environment state.

The research work in [1], shows RL applied to uplink power control. Outer-Loop Power Control (OLPC) in 5G networks relies on tuning two primary parameters, the normalized transmit power density $P0$, and the path-loss (PL) compensation factor α_{pl}. Optimization of these parameters is known to be of great importance to reach high uplink performance. One approach is to optimize uplink power control via an RL agent for each cell, controlling both $P0$, and α_{pl} parameters within a single neural network rather than focusing on $P0$ alone. However, mitigation actions are needed to cope with behavior resulting from multi-agent RL, such as high-power consumption from uncoordinated competition among gNBs in the network trying to maximize their own performance. To mitigate these issues, cooperative time synchronized reward mechanisms and sharing of state information between nearby RL agents can be implemented. Hence, achieving a common goal across multiple gNBs.

The solution in [1] is based on Double Deep Q Network (DDQN), where soft updates take place at every training occasion. In this solution, the neural network's output layer is divided in two dimensions, one dedicated for $P0$, and the other for α_{pl} indices (See Fig. 2). 3GPP defines 114 values for $P0$ and 8 values for α_{pl} resulting in 912 possible combinations.

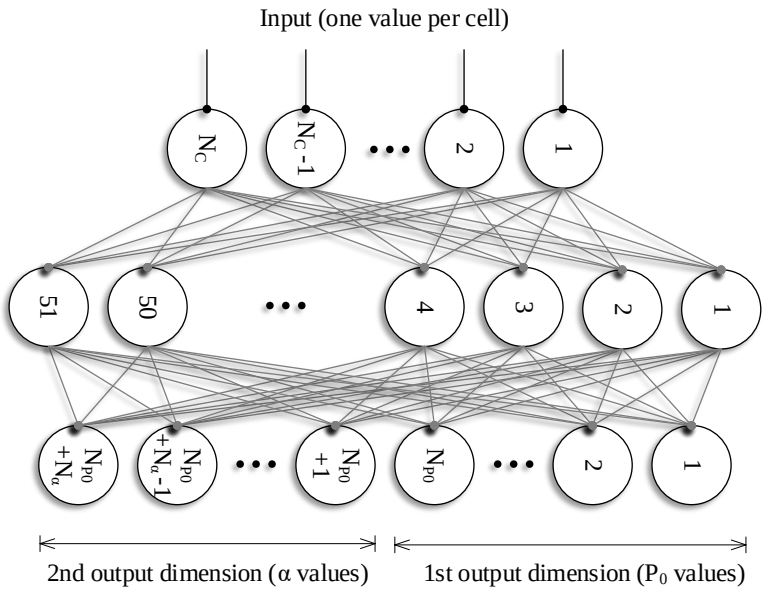


Fig. 2 Multi-Action Neural Network with Two Output Dimensions [1]

OLPC parametrization is the problem of finding a balance between the signal to interference plus noise ratio (SINR) and the number of resource blocks needed per transmission. If the gNB agents are only aware of their own parametrization and performance, such uncoordinated approach leads to a competition where the agents increase their transmission power to compensate the interference created by their neighboring agents. Hence, the agents should be provided with information that allows learning of power settings between gNBs, and that state information is shared between neighbors at each training step. Likewise, the reward is the sum throughput per utilized resource blocks over the closest neighbors including the agent's own cell.

The simulation result in [1] (See Fig. 3) shows that maximizing the neighborhood reward alone may result in unfair user and cell throughput, as power allocations can become widespread. Thus, an alternative is to carry out averaging of the ML-suggested actions, which yields a fairer and more uniform power allocation between cells. Similarly, co-operation is shown to be essential in multi-agent power control, as the co-operation range affects significantly the results. If the co-operation range is too high, it leads to noisy rewards which impairs learning, while without co-operation gains collapse and DDQN is unable to learn the full effects of its actions.

Additionally, when evaluated with the exhaustively searched best configuration common across all simulation realizations (referred as golden baseline), simulation results show that it is possible to achieve ~10% gain in cell throughputs in average, with the gain being rather fairly distributed over all UEs within the simulation, showing further benefit over traditional parametrization approaches.

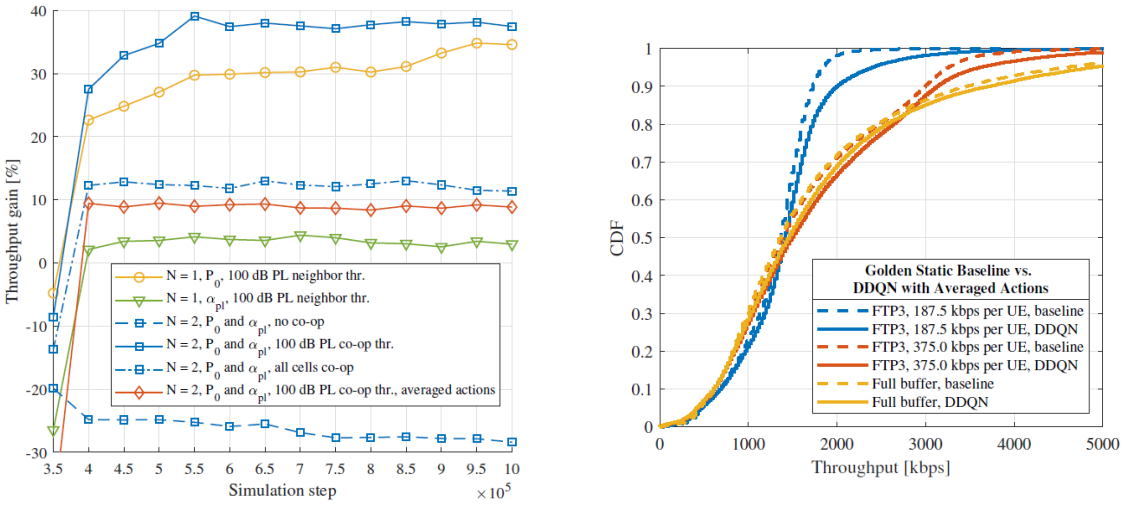


Fig. 3 Simulation results: (left) Average Cell Throughout Gain, where N=2 refers to learning both *P0* and α_{PL} output dimensions; (right) Windowed User Throughput Distribution of with different offered loads [1]

4. Conclusion

AI/ML-based solutions have the potential to further extend the boundaries of performance of the air interface. However, to deploy AI/ML solutions at scale, standardization of lifecycle management framework is needed. Hence, paving the way with work in 5G-Advanced for AI-native 6G, where AI/ML is considered from the start as a key design principle of the system.

Similarly, 6G development must specify enablers for more dynamic reconfiguration of system information parameters. Likewise, more dynamic power control as well as other machine learning applications, such reinforcement learning, bring performance beyond that of common parameters set over the network. Further, it could be expected that such machine learning algorithms will turn to be essential parts of 6G making the paradigm shift towards self-sustained networks, where multiple dependent parameters and inter-connected features must be tuned simultaneously on the fly.

REFERENCE

[1] P.Kela, and T.Veijalainen, *"Cooperative Action Branching Deep Reinforcement Learning for Uplink Power Control"*, in 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), 2023.

[2] A.Maeder, *"AI/ML unleashes the full potential of 5G-Advanced"–(Nokia)* https://www.nokia.com/blog/aiml-unleashes-the-full-potential-of-5g-advanced

[3] A.Maeder, and I.Kovacs, *"Scaling up AI/ML for cellular radio access"–(Nokia)* Scaling AI/ML for cellular radio access in 5G-Advanced (nokia.com)

AI/ML-based Radio Propagation Prediction Technology

Tetsuro Imai, Tokyo Denki University
Koshiro Kitao, NTT DOCOMO. INC.
Satoshi Suyama, NTT DOCOMO. INC.

***Abstract*—Recently, advancement of AI/ML has been remarkable, and many applied research studies are attracting attention now. This is also true in the field of radio propagation. This paper introduces its application to radio propagation prediction, which is currently under intensive study.**

1. Introduction

In recent years, artificial intelligence (AI) / machine learning (ML) has made remarkable progress, and many applied research studies have been reported. Here, they are mainly based on deep learning. The deep learning is one of the methods of ML for neural networks with many layers (or DNN: deep neural network). Deep learning has succeeded the dramatic performance improvement of image recognition, natural language processing etc., while utilizing of abundant computer resources and big data. The main reason for its success is that the deep learning can automatically extract features of contents.

In mobile communications, accurate prediction of radio propagation characteristics is needed for optimum cell design, various prediction models have been proposed so far [1]. These are categorized into two types. One is physical-based model which is based on electromagnetic theory, and another is statistical (or data-driven) model which is based on measurement data. Here, ray tracing (RT) is one of the physical-based models and has become popular tool for radio propagation analysis in recent years. In RT, various propagation characteristics such as loss, time of arrival, angle of arrival and so on can be predicted by tracing rays between transmitter (Tx) to receiver (Rx) while taking interaction (reflection, diffraction, transmission) into account. However, increasing the number of interactions considered to improve the prediction accuracy increases the computation time. So, when the target characteristic is only propagation loss, the statistical model, e.g. Okumura-Hata model [2] is preferred.

In statistical modeling, multi-regression analysis has been applied to model the data [3]. The multi-regression analysis is a very powerful tool, but it is needed to manually determine input parameters (especially environmental parameters related to building, street, etc.) and functional form beforehand. This is very difficult because there are a lot of candidates. So, the prediction models with neural network (NN) have been proposed in [4], [5]. By using these models, functional form is automatically generated,

and it is reported that prediction accuracy for propagation loss is improved. However, the models are based on conventional fully connected neural network (FNN), optimal input parameters must be investigated, manually.

As mentioned above, the deep learning can automatically extract features of contents. Especially, deep convolutional neural network (DCNN) are very useful to extract features from image. This means that optimal parameters for propagation loss prediction can be automatically obtained from map data with information such as building spatial distribution. So, DCNN-based model has been proposed for propagation loss prediction [6] and is currently being vigorously studied [7]-[12]. This paper presents our latest results in [12].

2. DCNN-based Radio Propagation Prediction Model

2.1 DCNN Configuration

DCNN of our proposed model is constructed by two parts: feature extraction part and prediction part, as show in Fig. 1.

The feature extraction part is for extraction of features of contents as key parameters for propagation loss prediction, and it is constructed by DCNN which has 13 convolutional layers: Conv_1 – Conv_13, and five max. pooling layers: Pool_1 – Pool_5. First, three maps (the size of each map: 256-by-256) are input. In Conv_1&2 layers, convolutional processing with 32 filters (the size of each filter: 3-by-3) is done and then the 32 maps (the size of each map: 256-by-256) are obtained. In next Pool_1 layer, max. pooling processing is done for 32 maps. Here, pooling size is 2-by-2, so the size of output map is reduced to 128-by-128. After the similar convolutional and pooling processing are repeated, 256 maps (the size of each map: 8-by-8) are output from Pool_5 layer. Here, the number of samples is 16384 (=8×8×256) and these are input to Dense_1 layer after conversion process to 1 D data in Flatten_1. The prediction part is constructed by FNN with two fully connected layers: Dense_1 and Dense_2. After the processing in Dense_1&2, propagation loss is predicted as output. Note that activation function is defined as: $f(x) = x$ in Dense_2 layer; otherwise, Rectified Linear Unit function, i.e. $f(x) = \max(0, x)$.

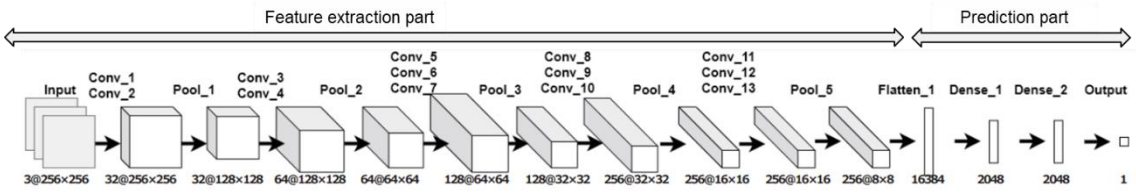


Fig. 1 DCNN configuration

2.2 Input Map Data

In our model, the spatial information of rectangular area centered on mobile station (MS) position is input to DCNN as map data. The size of rectangular is 256 m -by- 256 m, and the area is sampled with 1 m mesh, so, the sample size is 256-by-256. In addition, the rectangular is defined so that the base station (BS) always exist in a certain direction. Specifically, as shown in Fig. 2, the rectangular region is defined so that BS is oriented positively on the x_m axis in the local coordinates of the map with MS as the origin. By this definition, the spatial information about "BS direction" are indirectly considered for DCNN learning, even if the BS position are not directly input to the DCNN as parameter.

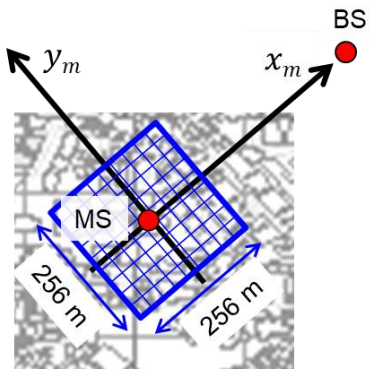
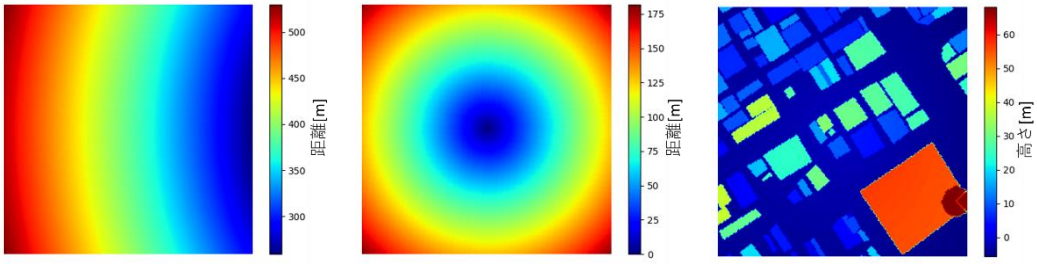


Fig. 2. Definition of rectangular region

Input maps are three as follows.

- BS distance map: Map with distance from BS to each mesh as an element.
- MS distance map: Map with distance from MS to each mesh as an element.
- Building map: Map with building height information in each mesh.

In the building map, the height is normalized by the height of Fresnel-zone center when assuming one time scattering. This advantage is that BS antenna height and MS antenna height are indirectly considered as input parameters. Figure 3 shows the examples of input map data.



(a) BS distance map (b) MS distance map (c) Building map

Fig. 3. Examples of input map data

3. Performance of DCNN-based Model

3.1 Measurement Data

Propagation loss data measured in Kokura area are used for performance evaluation. Here, the data can be obtained for free from AP propagation database [13]. Figure 4 and Table 1 show the measurement area and conditions, respectively.

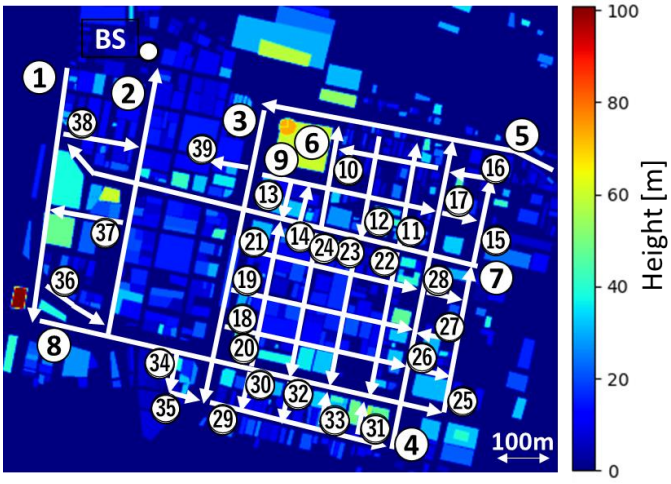


Fig. 4 Measurement area (Kyushu Kokura area, Japan): White lines represent measurement courses.

Table 1 Measurement conditions

Frequency	1298 MHz
Transmission power	39.5 dBm
BS antenna	$\lambda/2$ dipole antenna (2dBi)
MS antenna	
BS antenna height	12.5 m
MS antenna height	1.5 m

In this paper, the data of 5 courses (#6, #19, #24, #27, #32) are used for validation, the remaining data of 29 courses are for DCNN training. Here, data of course #5 is not used because sufficient input map data could be obtained. The total number of samples (or MS points) is 81 for validation and 713 for training.

3.2 Evaluation Results

Figure 5 shows the prediction results for validation data. Horizontal axis represents distance from BS and vertical axis represents propagation loss. We find that measurement and prediction are agree well. Here, RMS error is 3.23 dB.

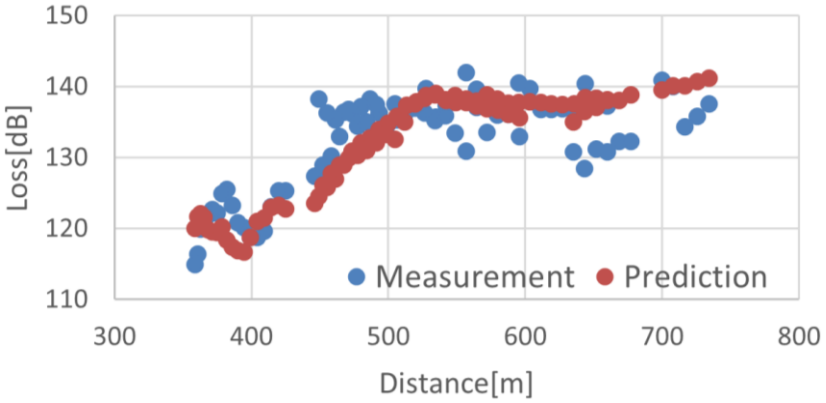
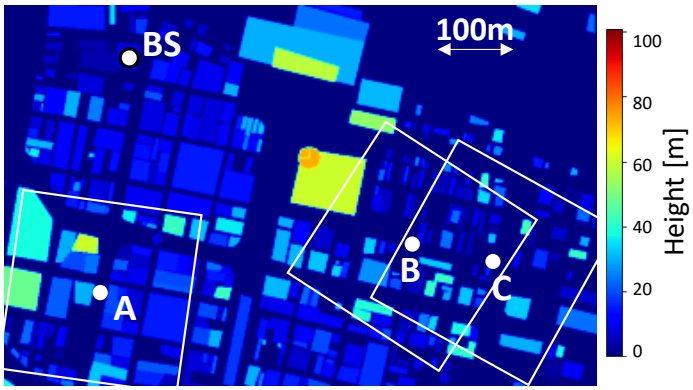
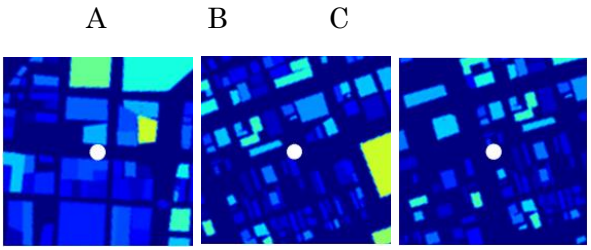


Fig. 5 Prediction results.

The extracted features after training DCNN can be visualized by using Grad-CAM (Gradient-weighted Class Activation Mapping) [14], which one of XAI (Explainable AI) algorithms. Therefore, Grad-CAM were performed for three points as shown in Fig. 6. Figure 7 shows the analysis results with Grad-CAM. In Fig. 7, the larger the gradient value, the higher the contribution for the propagation loss prediction. From the results, DCNN-based model is thought to use the "distribution of low-rise buildings and spaces without buildings" in the vicinity of MS as the basis for determining the propagation loss prediction.

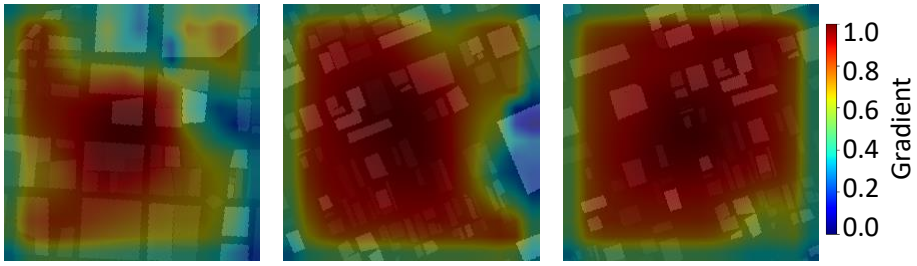


(a) Positional relationship with BS



(b) Maps in local coordinate system

Fig. 6 Reception points for evaluation of extracted features from map data.



(a) Point A (b) Point B (c) Point C

Fig. 7 Analysis results with Grad-CAM when using multiple maps.

Finally, Fig. 8 shows propagation loss distribution predicted by trained DCNN when BS are installed in different location. Note that the other propagation conditions are same as that in table I. From this figure, we can see that even if the distance from the BS is the same, the propagation loss increases in areas with dense buildings.

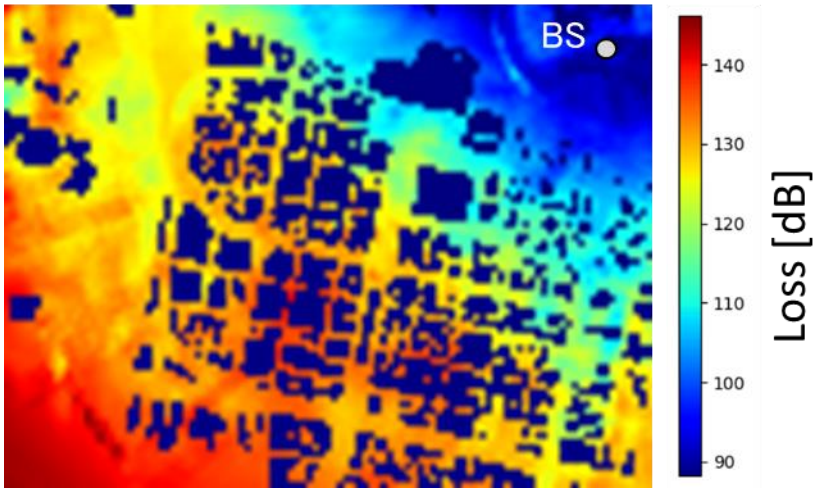


Fig. 8 Propagation loss distribution predicted by trained DCNN.

4. Conclusion

In this paper, we introduced DCNN-based model for radio propagation loss prediction. This model predicts the propagation loss from map data with information such as building spatial distribution and its prediction accuracy is higher than conventional model based on multi-regression analysis. In our study, RMS error of about 3 dB is obtained. And also, we showed that the basis for determining the prediction in the DCNN-based model can be confirmed by Grad-CAM.

REFERENCE

[1] T. K. Sarkar, Z. Ji, K. Kim, A. Medour, and M. Salazar-Palma, “A Survey of Various Propagation Models for Mobile Communication,” IEEE AP Magazine, Vol. 45, No. 3, pp, 51-82, June 2003.
[2] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” IEEE Trans. VT, vol. 29, no. 3, pp. 317-325, Aug. 1980.
[3] K. Kitao, and S. Ichitsubo, “Path loss prediction formula in urban area for the fourth-generation mobile communication systems,” IEICE Trans. Commun., vol. E91-B, no. 6, pp. 1999-2009, June 2008.
[4] E. Östlin, H. Zepernick, H. Suzuki, “Macrocell Path-Loss Prediction Using Artificial Neural Networks,” IEEE Trans. VT, vol. 59, no. 6, pp. 2735-2747, July 2010.
[5] M. Ayadi, A. Ben Zineb, and S. Tabbane, “A UHF Path Loss Model Using Learning Machine for Heterogeneous Networks,” IEEE Trans. AP, vol. 65, no. 7, pp. 3675-3683, July 2017.
[6] T. Imai, K. Kitao, and M. Inomata, "Radio Propagation Prediction Model Using Convolutional Neural Networks by Deep Learning," EuCAP2019, April 2019.
[7] T. Hayashi, T. Nagao, and S. Ito, “A study on the variety and size of input data for radio propagation prediction using a deep neural network,” EuCAP2020, March 2020.
[8] N. Kuno, W. Yamada, M. Inomata, M. Sasaki, Y. Asai, and Y. Takatori, “Evaluation of Characteristics for NN and CNN in Path Loss Prediction,” ISAP2020, Jan. 2021.
[9] X. Zhang, X. Shu, B. Zhang, J. Ren, L. Zhou, and X. Chen, "Cellular Network Radio Propagation Modeling with Deep Convolutional Neural Networks," in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 2378-2386, Aug. 2020.
[10] T. Nagao, T. Hayashi, “A Study on Urban Structure Map Extraction for Radio Propagation Prediction using XGBoost,” EuCAP2021, March 2021.
[11] K, Inoue, K. Ichige, T. Nagao, and T. Hayashi, “Learning-Based Prediction Method for Radio Wave Propagation Using Images of Building Maps,” IEEE AWPL, vol. 21, no. 1, pp. 124-128, Jan. 2022.
[12] K. Kozera, T. Imai, K. Kitao, and S. Suyama, “Performance Evaluation of DCNN-Based Model for Radio Propagation Loss Prediction - Analysis on Prediction Mechanism with Grad-CAM -,” IEICE Trans. Commun., vol. J106-B, no.9, pp. 618-627, Sep. 2023.
[13] AP Propagation Database: Online data repository created and supported by Technical committee on Antennas and Propagation, IEICE.
https://www.ieice.org/cs/ap/language/en/misc-eng/denpan-db/
[14] R. R. Selvaraju, et al.,” Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” International Journal of Computer Vision, Dec. 2019.

6G Network AI Architecture for Everyone-Centric Customized Services

Huang Huanhuan, Huawei Technologies
Peng Chenghui, Huawei Technologies
Yang Yang, The Hong Kong University of Science and Technology
Koshimizu Takashi, Huawei Technologies Japan

Abstract— **This article proposes a network Artificial Intelligence (AI) architecture with integrated network resources and pervasive AI capabilities for supporting customized services with guaranteed QoEs (Quality of Experiences). Extensive simulations show that the proposed network AI architecture can consistently offer a higher User Satisfaction Ratio (USR) performance than the cloud AI and edge AI architectures with respect to different task scheduling algorithms, random service requirements, and dynamic network conditions.**

1. Three AI Architectures

1.1 Cloud AI

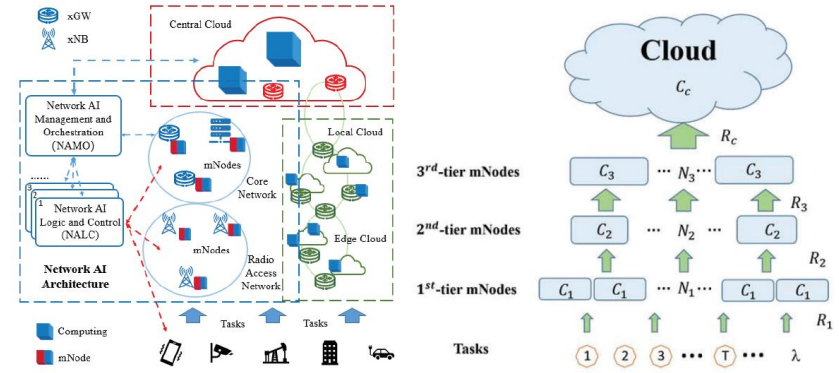
In the era of 5G, the cloud AI architecture has been widely adopted to provide centralized computing services. The conventional "cloud-pipe-terminal" structure decouples the data sensing functions at user terminals, the communication functions in mobile networks (a.k.a. the pipe), and the computing functions or the AI-enabled analytical services on the cloud [1]. This is simply a combination of the existing infrastructures of Data Technology (DT), Communication Technology (CT), and Information Technology (IT). It is very challenging to coordinate these separate functions in multiple facilities for effectively providing an agile, smooth, and stable service with guaranteed QoE.

1.2 Edge AI

In order to solve the problem of low speed, long delay, poor privacy, and high carbon emissions in centralized AI applications on the cloud, the edge AI architecture extends the computing capability from the cloud to the locations physically closer to end users. The edge AI architecture is much more effective in supporting computing-intensive, delay-constrained, security-assured, and privacy-sensitive applications, such as interactive Virtual Reality/Augmented Reality (VR/AR) games, but the costs for deploying edge clouds (also called cloudlets) widely in the neighborhood are usually very high.

As shown in Fig. 2-1 (a), central, local, and edge clouds are connected by high-speed, expensive bearer networks, which are just the traffic pipes with huge bandwidth. Local

and edge clouds are deployed as affiliated Over-The-Top (OTT) services to support computing-intensive applications. They are usually co-located with the existing network elements, but not embedded in mobile networks. Thus, cross-domain resource coordination and service orchestration between these local/edge clouds and end users require round-trip data transmissions through the mobile network. The actual service procedure is very complicated, time-consuming, and expensive, and may generate a series of management and technical problems such as redundant deployment costs, circuitous data paths, and frequent desynchronized cooperation. It is very difficult for the cloud AI and edge AI architectures to guarantee E2E (End to End) QoE for sophisticated cross-domain services in dynamic application scenarios and mobile environments.



(a) Deployments of Cloud, Edge, and Network AI Architectures.

(b) System Model.

Fig. 1 Three AI Architectures and the System Model.

1.3 Network AI

To address those challenging problems, this article proposes the network AI architecture with multi-tier multi-function Nodes (mNodes) to integrate and coordinate cross-domain Sensing, Storage, Communication, Computing, Control, and AI (S^2C^3A) resources for processing local/regional user data, executing distributed AI algorithms, and providing customized services for everyone as closely as possible. This architecture shifts the classic design paradigm that assumes mobile networks only as the pipe for data transmissions. Based on the hierarchy of mNodes, heterogeneous network resources and separate functions are effectively integrated to support cross-domain, wide-area, and delay-sensitive applications.

In Fig. 1 (a), the proposed network AI architecture consists of three key units and constructs a comprehensive, distributed, and scalable AI as a Service (AIaaS) platform in 6G. First, the network infrastructure is composed of dispersive mNodes in multi-tier

mobile networks. An mNode not only coordinates local resources as a Service Provider does for E2E service auction [2], but also integrate the basic S^2C^3A resources and multiple functions to support QoE-guaranteed, everyone-centric customized services. Besides general-purpose computing units, it is envisaged that more and more AI processors will be widely integrated and shared by the mNodes to provide the 6G native AI service platform. Second, each Network AI Logic and Control (NALC) unit is task-oriented and manages the multi-tier mNodes in a specific local/regional area through effective signaling schemes. In 6G mobile networks, a NALC coordinates the integrated S^2C^3A resources and functions for serving every task in real-time and near-real-time applications, i.e., E2E delay ranges from milliseconds to tens of milliseconds. The customized service procedure and personal QoE of every task are constantly monitored and optimized by a corresponding NALC. Third, a Network AI Management and Orchestration (NAMO) unit manages the AIaaS platform with multiple NALCs to support wide-area applications by cross-domain resource coordination, service orchestration, and E2E QoE guaranteeing protocols. The proposed network AI architecture can either serve various tasks independently, or complement with the cloud AI and edge AI architectures to satisfy sophisticated user requirements with challenging SRZ targets.

2. System Model and Simulation Results

2.1 System Model

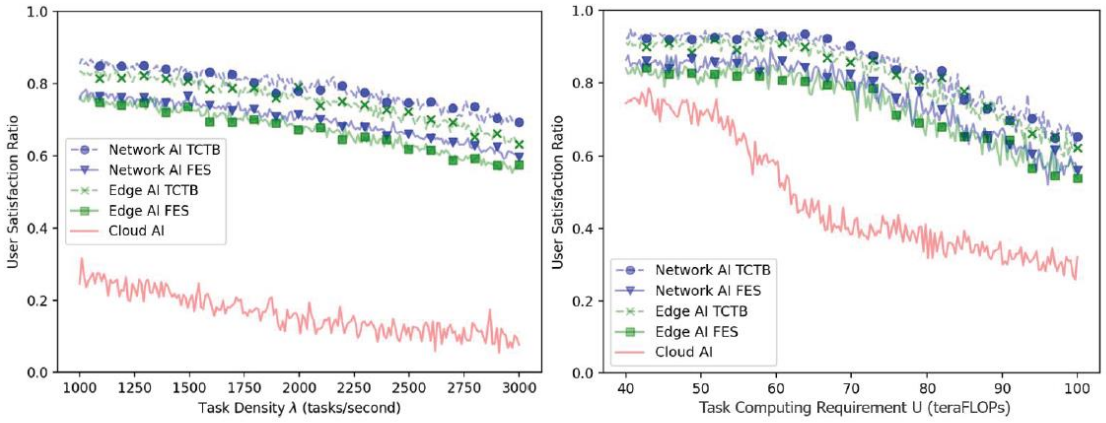
To study a typical 6G system with dispersive computing resources and pervasive intelligence, Fig. 2-1 (b) shows a general system model for different AI architectures. Let us consider a series of tasks, each having a customized Service Requirement Zone (SRZ), arriving at the system with a predefined rate. These tasks are generated randomly either by end users enjoying mobile internet services or by various devices and things embedded in industrial IoT applications. We consider a three-tier network AI architecture with three types of mNodes, which are represented by blue rectangular boxes. Above them sits a cloud, which has the highest data rate and the strongest computing power. This system model can be easily simplified to represent the cloud AI and edge AI architectures by setting the number of mNodes to zero from the first and second tier, respectively.

For an arbitrary task, the corresponding service provisioning procedure is determined by the specific task scheduling algorithm. Upon the arrival of task, its SRZ is first checked by a nearby 1^{st}-tier mNode at the edge, which analyzes the possibility of satisfying that SRZ with the network resources available in the vicinity. If local resources are sufficient, the task will be immediately served by this mNode. If not, a more powerful 2^{nd}-tier mNode will be initiated to lead the effort of identifying feasible

network resources in a bigger neighborhood. If regional resources are still not sufficient, an even stronger 3^{rd}-tier mNode will be called upon to perform multi-domain resource coordination over a much wider area.

2.2 Simulation Results

The simulation study is conducted in this article to compare the performance of three AI architectures, where two task scheduling algorithms are considered in performance evaluation. The Fair Equal Scheduling (FES) algorithm assigns all the tasks in a random manner, with half going to the edge and half to the cloud for services. The-Closer-The-Better (TCTB) algorithm follows the Pareto principle, or the 80/20 rule, so that 80% and 20% of all the tasks go to the edge and the cloud, respectively.



(a) Impact of Task Density when Task Computing Requirement U~N (70,1).

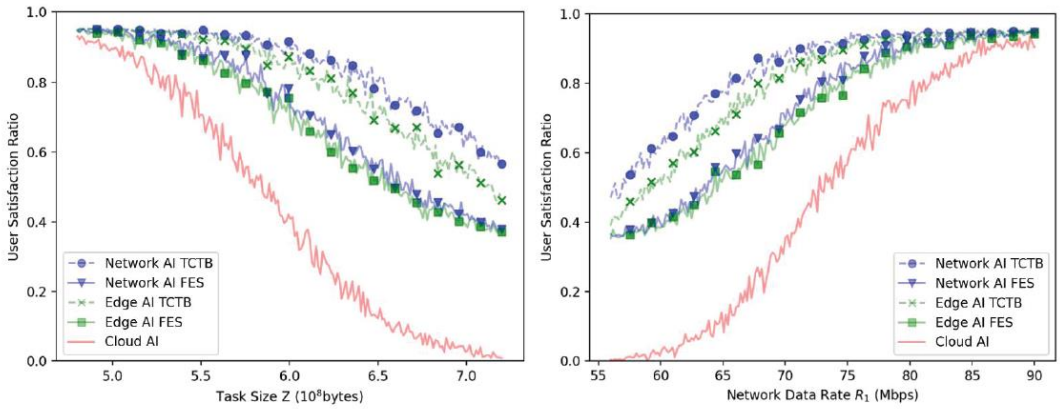
(b) Impact of Task Computing Requirement when Task Density λ=1000.

Fig. 2 USR versus Task Density and Computing Requirement.

Fig. 2 illustrates the User Satisfaction Ratio (USR) performance of the three AI architectures under dynamic task densities and computing requirements, where the USR is calculated as the ratio between the number of satisfied tasks and the total number of served tasks. In Fig. 2 (a), the task density has a linear impact on the decline of the USR curves under different AI architectures. And the network AI architecture achieves higher USR than the edge AI architecture, and much higher USR than the cloud AI architecture. In Fig. 2 (b), the USR curve of the cloud AI architecture has two knee points at about task computing requirement U=48 teraFLOPS and U=66 teraFLOPS. The transition region between them has a steep slope, which implies that the energy consumptions for executing all the tasks in the cloud increase very rapidly when the average computing requirement increases. Under both TCTB and FES

algorithms, the green and blue curves of the edge AI and network AI architectures are much less sensitive to this change, which is due to the efficient services by mNodes in the neighborhood.

In Fig. 3 (a), when task size increases, the USR curve of the cloud AI architecture degrades dramatically because long-distance transmissions of bigger tasks become more time-consuming and energy-intensive, thus adversely impacting the USR. On the contrary, the USR curves of the edge AI and network AI architectures are much less sensitive to task size changes, thanks to the computing resources deployed at the edge and in the network. Fig. 3 (b) demonstrates the influence of network data rates on the USR performance. These curves are like the mirror flips of those in Fig. 3 (a), because higher network data rates and smaller task sizes both imply lower transmission delays. Therefore, increasing network data rates and reducing task size have almost equivalent impact on the USR performance.



(a) Impact of Task Size when Network Data Rate R_1~N (70, 7).

(b) Impact of Network Data Rate when Task Size Z~N $(6 \times 10^8, 10^6)$.

Fig. 3 USR versus Task Size and Network Data Rate.

3. Conclusion

The cloud, edge, and network AI architectures were studied and compared in this article under dynamic task densities, task sizes, computing requirements, network data rates, and two task scheduling algorithms. By deploying multi-tier mNodes, the proposed network AI architecture with integrated S^2C^3A resources can effectively support customized services for a variety of user tasks, thus achieving the highest USR under random service requirements and dynamic network conditions. In contrast, the centralized cloud AI architecture has difficulties in meeting stringent delay and energy consumption bounds, thus not suitable for delay-sensitive broadband applications.

REFERENCE

[1] N. Chen, Y. Yang, T. Zhang, M. T. Zhou, X. L. Luo, and J. Zao,“Fog as a Service Technology,” IEEE Communications Magazine, Vol. 56, No. 11, pp. 95-101, Nov. 2018.
[2] X. Chen, Y. Deng, G. Zhu, D. Wang, and Y. Fang, “From Resource Auction to Service Auction: An Auction Paradigm Shift in Wireless Networks,” IEEE Wireless Communications, early access, May 2022.

AI-based Application-aware RAN Optimization

Eiji Takahashi, NEC
Takeo Onishi, NEC
Yoshiaki Nishikawa, NEC

Abstract— **It has become increasingly important for industries to promote digital transformation (DX) by utilizing Beyond 5G, Industrial Internet of Things (IIoT), and Artificial Intelligence (AI) to realize a highly productive and prosperous society. In addition to conventional policies of improving the average quality of experience (QoE) at each mobile coverage area, there is an increasing need to strengthen policies that precisely adhere to quality of service (QoS) requirements per communication session and in real-time to enable the stable use of applications at high-performance levels, e.g., work speed or productivity. This article introduces an AI-based application-aware radio access network (RAN) optimization technology that can support such policies based on Open RAN architecture.**

1. Introduction

Because of the labor shortage and the decrease in skilled workers due to the declining birthrate and aging population, there is an increasing need to replace humans with machines in several tasks to solve social issues. Accordingly, there is a need for automation, remote monitoring/control, and labor-saving by promoting digital transformation through the utilization of Beyond 5G, IIoT, and AI to realize a highly productive and prosperous society [1]. In the IIoT area, there are many use cases that require the mobility and ease of equipment installation, and where wireless communication is essential. And high IIoT application performance (working speed, productivity, etc.) must be ensured, often resulting in stricter requirements for QoS.

Thus, in addition to the policy to improve the average QoE of users at each area, there is a need to strengthen policies that precisely protect IIoT application performance (work speed, productivity, etc.) per communication session. For this purpose, it is necessary to intelligently and autonomously control the RAN according to the conditions of application, network, and site, so that IIoT applications can be used stably at high performance. This article introduces an AI-based application-aware RAN optimization technology that can support such policies based on the Open RAN architecture [2], [3].

2. Application-aware RAN Optimization

Here, the industrial use case to ensure the uninterrupted transport of materials at the factory or warehouse floor by means of Automated Guided Vehicles (AGVs) is considered. Communication services for those remote-control applications need to fulfill stringent requirements, especially in terms of latency, communication service availability and determinism. In those applications, two-way communication consisting of robot status monitoring and control instructions must be completed at a constant cycle, and the system safely stopped with fail-safe if the latency exceeds a threshold. Safety is maintained by fail-safes, but if fail-safes occur frequently, the facility utilization rate and productivity will decrease.

The application-aware RAN optimization method consists of AI that analyzes communication requirements and radio quality fluctuations on a per-user terminal basis, such as robots and vehicles, and AI that dynamically controls RAN parameters on a per-user terminal basis based on the results of that analysis. This AI learns from past operational records of robots and vehicles, and optimally controls RAN parameters such as modulation and coding scheme (target block error rate), radio resource allocation (resource block ratio), and maximum allowable delay (delay budget) while predicting the probability of exceeding communication latency requirements.

In a typical 5G network, RAN parameters are fixed and set for the entire network. However, this technology dynamically controls them on a per-user terminal basis to improve application productivity. Based on the architecture shown in Fig. 1, the proposed method can perform the following tasks on a per user equipment (UE) basis in near-real-time: 1) estimating application QoS requirements based on information supplied by the application server, 2) predicting fluctuations in wireless quality by using radio quality information supplied by the central unit (CU) and distributed unit (DU), and 3) proactively optimizing CU/DU parameters [4]. While running machine learning, the system verifies that the accuracy is not compromised. If a risk is detected, it switches to a stable logic-based engine. This technology ensures the stability of RAN control by switching engines.

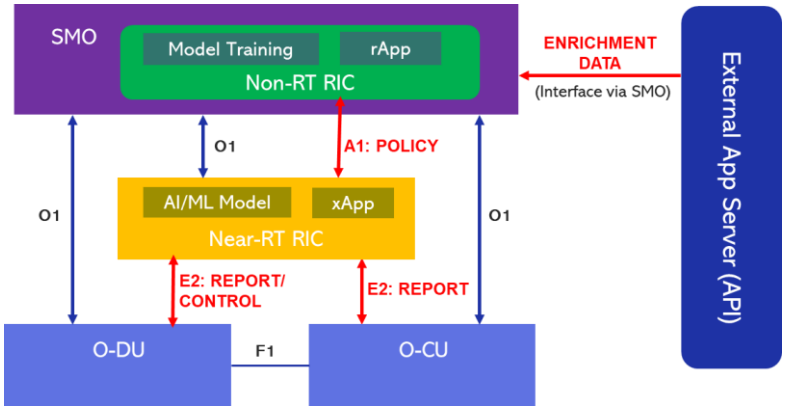


Fig. 1 Architecture

3. Evaluation

By using mobility, radio propagation, and network simulators, the timeliness and availability of communication services are evaluated in the context of mean time between failures, which is one of the important key performance indicators (KPIs) defined by 3GPP for this type of traffic. The proposed method optimizes RAN parameters, including modulation and coding scheme (target block error rate), radio resource allocation (resource block ratio), and maximum allowable delay (delay budget), on a per-UE basis in near real-time. The proposed method is compared with the method of using typical settings for these RAN parameters in conventional methods. Table 1 showcases the simulation conditions, while Fig. 2 illustrates the simulation results in an environment where the QoS requirements and radio quality vary based on field conditions (e.g., movement of UEs, distance between UEs, etc.).

Table 1 Simulation Conditions

Number of gNB and cell	gNB: 1
	cell: 1
Frequency	4.8GHz (n79)
Bandwidth	100MHz
SCS	30kHz
Duplex	TDD
DL, UL ratio	DL:UL = 1:1
Transmission power	23dBm
Floor area	100m x 100m
Floor layout	layout assuming a factory
Number of robots	18 in maximum
Robot running speed	3 m/s in maximum
Traffic per robot	DL: 150Kbps in maximum UL: 1 Mbps in maximum

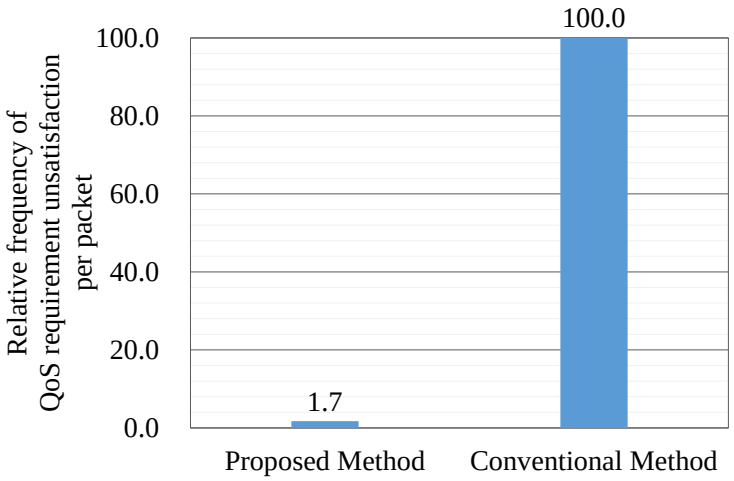


Fig. 2 Simulation Results

The simulation results of implementing this technology to a system that remotely controls multiple autonomous robots operating in factories or warehouses have confirmed that the proposed method reliably satisfies the QoS requirements, in comparison to the conventional method. In other words, the number of robot stoppages can be reduced by 98% or more compared to cases where this technology is not used.

4. Conclusion

Mobile network specifications will become more sophisticated toward the Beyond 5G era, however, intelligent network optimization during operation will be crucial in responding to the changing conditions of applications, networks, and sites. To guarantee strict QoS requirements in the vertical domain and to support diversification of application requirements and wireless quality variation, an application-aware RAN optimization technology was proposed. The simulation results of applying this technology to a system that remotely controls multiple autonomous robots operating in factories or warehouses confirmed that the number of robot stoppages can be reduced by 98% or more compared to cases where this technology is not used.

Conventionally, IIoT devices are equipped with intelligent functions that are specific to its vendor or model, and the IIoT controller software is also tied to a specific vendor and model. When Beyond 5G realizes a reliable wireless communication environment with low latency, intelligent and high-load data processing will be possible on the cloud or edge server. This makes it easier for the IIoT controller installed in the cloud or edge server to coordinately control IIoT devices of multiple vendors, and for multiple models to optimize the entire system. Furthermore, achieving simplification, lightweight implementation, and generalization of IIoT devices will likely drive the spread of IIoT

solutions, and as a result, accelerate the developments in IIoT applications and AI systems.

Acknowledgements

This article is based on results obtained from "Research and Development Project of the Enhanced Infrastructures for Post-5G Information and Communication Systems" (JPNP20017), commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCE

[1] 3GPP, “Service requirements for cyber-physical control applications in vertical domains,” TS 22.104, V17.4.0, Oct 2020.

[2] NEC, “NEC develops RAN autonomous optimization technology that dynamically controls 5G networks based on user terminal status ~Remote control of robots and vehicles with high productivity~,” Feb. 16, 2024. https://www.nec.com/en/press/202402/global_20240216_01.html

[3] NEC, “More freedom in DX and advanced application development, Autonomous optimization of 5G networks with AI,” Feb. 16, 2024. https://www.nec.com/en/global/rd/technologies/202315/index.html

[4] ORAN Working Group 2, "AI/ML workflow description and requirements," Technical Report v1.1, 2020.