Beyond 5G White Paper
6G Radio Technology Project
“AI/ML and Digital Twin Technologies”

Version 1.0

May 7, 2025

XG Mobile Promotion Forum

XGMF

【Revision History】

Ver.	Date	Contents	Note
1.0	2025.5.7	Initial version	

Preface

Towards the Beyond 5G/6G era, the technological evolution of communication networks is progressing rapidly, and artificial intelligence (AI) and machine learning (ML) technologies will play a significant role in this evolution. The introduction of AI technology into 5G is already being considered at a rapid pace, and applications using AI are being installed in smartphones.

In addition, in the 6G era, digital twin (DT) technology is considered to be very important, in which the real world is reproduced in cyberspace, and data collected from the real world is used to simulate and emulate beyond the constraints of the real world using AI, etc., to gain new knowledge, and to feedback and utilize those knowledge to the real world.

These AI/ML and DT technologies will be used in various fields to enhance the capabilities of Beyond 5G/6G. Beyond 5G White Paper Supplementary Volume "AI/ML Technologies" already published by XGMF introduced (i) AI/ML technologies for network operation and management, (ii) AI/ML technologies for optimizing radio access resource management, and (iii) AI/ML technologies for user/application-centric communications.

This white paper targets the utilization of DT technology in addition to the AI/ML technologies mentioned in the above Beyond 5G White Paper Supplementary Volume. As 6G Radio Technology Project of XGMF, this white paper summarizes the latest trends and research and development (R&D) activities of the utilization of AI/ML and DT technologies in the 6G radio technology field. Specifically, it describes the trends of standardization in 3GPP and O-RAN toward 6G, as well as the status of global technology studies. In addition to the content of the previous Beyond 5G White Paper Supplementary Volume, this white paper also introduces in detail cutting-edge R&D efforts on the AI/ML and DT technologies for Beyond 5G/6G in Japan.

In conclusion, as technological evolution progresses towards Beyond 5G/6G, the AI/ML and DT technologies are emerging as one of key elements in this technological evolution. Japan is at the forefront of the next-generation mobile communications revolution due to its efforts to overcome challenges to realize the AI/ML and DT technologies and its commitment to R&D in this area. The purpose of this white paper is to provide a comprehensive overview of the potential, challenges, and future directions of the AI/ML and DT technologies for Beyond 5G/6G, with particular emphasis on initiatives and progress in Japan.

This white paper was prepared with the generous support of many people who participated in the AI/ML and DT Working Group of 6G Radio Technology Project, XGMF. The cooperation of telecommunications industry players as well as academia experts has also been substantial. Thanks to everyone's participation and support, this white paper was able to provide a lot of useful information for future discussions on business creation between 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.

Tomoaki Otsuki, Keio University

Satoshi Suyama, NTT DOCOMO, INC.

I. Trends of AI/ML and Digital Twin towards 6G

The rapid evolution of wireless communication technologies is paving the way for the next generation of connectivity, known as Beyond 5G or 6G. As we move towards this new era, the integration of artificial intelligence (AI) / machine learning (ML) and digital twin technologies becomes crucial. These technologies promise to revolutionize the way we design, manage, and optimize wireless communication systems, leading to unprecedented levels of efficiency, reliability, and user experience. This section introduces both standardization and technology trends of AI/ML and digital twin.

I-1. Standardization in 3GPP and O-RAN for AI/ML

Tetsuya Yamamoto, Hidetoshi Suzuki
Panasonic Holdings Corporation
Liqing Liu, Kozue Hirata
Sharp Corporation
Noboru Osawa, Yu Tsukamoto
KDDI Research, Inc.

The AI/ML technologies have been advanced remarkably in recent years. Applying them to wireless communications has received increasing attention. The integration of AI/ML to enhance network performance, efficiency, and scalability are active discussion in the standardization bodies such as 3rd Generation Partnership Project (3GPP) and Open Radio Access Network (O-RAN) alliance [1, 2]. This section provides a brief overview of the current AI/ML standardization status within 3GPP and O-RAN.

I-1.1. Standardization in 3GPP

In 3GPP, the initial specification of 5G is Release 15. Functional improvements and additions for 5G are Release 16 and 17. Release 18 and later are named as 5G-Advanced by 3GPP. In 5G-Advanced, the system performance and efficiency of enhanced mobile broadband (eMBB) are improved to address the short-term needs. Additionally, the use cases and services are further expanded to address the need of various verticals like satellite industry. Furthermore, new technology domains based on long- to medium-term needs like application of AI/ML are studied and standardized [3]. 3GPP will start discussions on 6G from Release 20 and to standardize it in Release 21. A workshop on 6G was held in March 2025.

AI/ML can be applied to various purposes. Therefore, numerous studies and standardization in areas such as applications, media, system management and radio access network (RAN), are progressing within 3GPP. This section describes the status related to application of AI/ML to RAN, which is specific to wireless communication. The application of AI/ML to RAN can be categorized into two types: 1) applying AI/ML technologies within the wireless network, such as base stations (BSs), and 2) applying AI/ML technologies to the air interface, which is the communication between user equipment (UE) and BS.

In the application of AI/ML within RAN, AI/ML technologies could be applied to various BS processing to achieve optimal network management and the network performance enhancement. In Release 17, the application of AI/ML in three areas were studied [4]: 1) network energy saving, 2) load balancing, and 3) mobility optimization. Standardization of them was completed in Release 18 in June 2024 [5, 6]. Furthermore, in Release 19, functional extensions for new use cases such as network slicing and optimization of coverage and capacity are ongoing. Additionally, items that not finalized in Release 18 are also under discussion, including the case that base stations functions are split into centralized nodes and distributed nodes [7]. For Release 20 5G-Advanced, new AI/ML-based use cases based on current 5G architecture and interface are considered with QoE optimization, network energy saving, and mobility (including multiple-hop target node UE trajectory) as the potential candidates for study [8].

In the application of AI/ML to air interface, AI/ML technologies could be utilized for enhanced and efficient performance. In Release 18, feasibility of the application of AI/ML to air interface was studied [9, 10]. This involved the general framework discussion for AI/ML for air interface, as well as the study of specific use cases such as the channel state information (CSI) feedback enhancement, beam management, and positioning. In Release 19, the actual standardization work is on the progress to realize use cases, such as beam management, positioning, and CSI prediction, where the introduction of AI/ML technologies would be effective, based on the study from Release 18 [11]. In addition, use cases that were not concluded in Release 18 (CSI compression) and new use case (mobility) are also to be studied [12, 13]. These use cases are the potential candidates for normative work in Release 20 5G-Advanced [8]. The following provides an overview of the study of AI/ML to air interface in Release 18 and the current discussion status in Release 19.

I-1.1.1. Framework of AI/ML to Air-interface

For the framework discussion of AI/ML to air interface, life cycle management (LCM), which is the process for appropriately utilizing AI/ML models, was addressed. It was identified that the framework for AI/ML application includes elements such as data collection, model training, model storage and transfer, inference using models, and model management. The relationships among these elements are summarized in Fig. I-1.1.1-1.

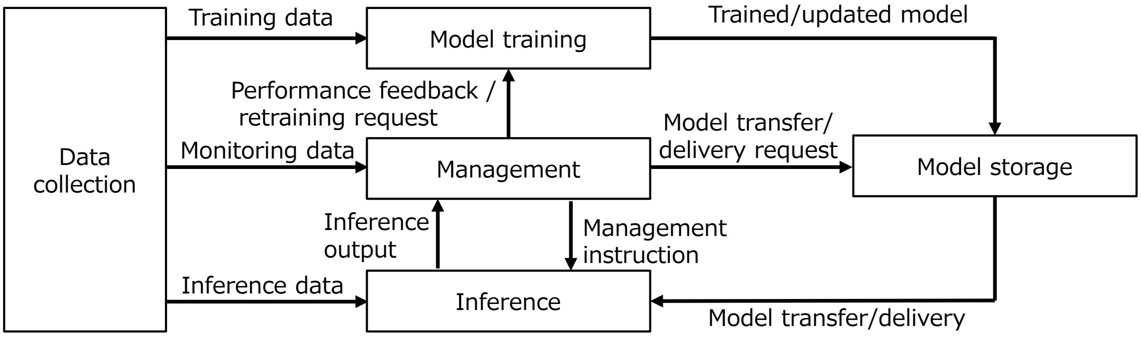


Fig. I-1.1.1-1: Framework / LCM for applying AI/ML

In Release 18, three levels of coordination were discussed.

- Level X: AI/ML is implemented, but there are no AI/ML-specific standardization extension. For example, AI/ML may be used for channel estimation at UE without informing the information related to the usage of AI/ML to the BS.
- Level Y: There is AI/ML-specific enhancement using the control signal between the UE and the network, but there is no transfer of AI/ML models. For instance, based on the radio conditions and BS configuration, the network may instruct the UE to perform specific AI/ML operations.
- Level Z: There is a transfer of AI/ML models between the UE and the network. For example, AI/ML models trained within the 3GPP network are transferred to the UE for the inference.

In Release 18, two different types of LCM were identified for controlling UE-side AI/ML models from the network.

- Functionality-based LCM
- Model-ID-based LCM

In functionality-based LCM, the network controls functionalities without aware of the UE's AI/ML models. By utilizing control signals, the network instructs the selection, activation, deactivation, switching of AI/ML-enabled functionalities, and fallback to non-AI/ML functionalities, while the actual management of the models are within the UE. In model-ID-based LCM, the network is aware of the UE's AI/ML models and controls them. Models are identified by model IDs and can be categorized into; 1) physical models, where the model structure and parameters are shared between UE and network, and 2) logical models, where the actual model structure and parameters are not shared but certain characteristics are shared between UE and network.

When AI/ML models are trained with real field data, they are influenced by not only the parameters defined by 3GPP but also conditions that include implementation-specific scenarios of networks and UEs. Examples of them are BS antenna beam shapes and beam control, power control, and implementation-specific receiver algorithms. These are called as additional conditions. There is ongoing discussion on which of these parameters and conditions must be aligned between the training and the inference for the efficient usage. Ideally, these additional conditions should be aligned between the training and inference. However, these conditions can contain proprietary information held by stakeholders such as network operators, network vendors, UE vendors, and users. Therefore, methods to align the additional conditions without disclosing proprietary information as much as possible are studied. One of examples is to exchange the trained model parameters and/or dataset between UE and network, which would not disclose proprietary information.

For the efficient usage of AI/ML-enabled functionalities in wireless communication, it is crucial that good performance can be maintained across various scenarios and conditions, such as different mobility speeds, radio propagation environments, and BS antenna configurations. Two main approaches have been discussed for achieving this: 1) model generalization, and 2) model switching. Model generalization means a single AI/ML model is generalized to handle different scenarios and varying BS antenna configurations by using diverse datasets. This approach may lead to larger model sizes and increased complexity, which can pose implementation complexity and power consumption on UEs. Model switching, on the other hand, entrails using AI/ML models that are tailored to specific conditions, such as particular cells. Then, network or UE selects appropriate AI/ML models for the specific conditions. While each model may be less complex and potentially offer higher performance, this approach has challenges to determine which model to be used in certain conditions or environments and how to manage and control larger number of models. For example, to select to appropriate model could require sharing proprietary network-side information with the UE, as previously mentioned.

I-1.1.2. CSI Feedback Enhancement

Accurate CSI is vital for optimal link adaptation and resource allocation. In 3GPP, CSI feedback mechanism involves the UE measurements and CSI reporting. However, a temporal delay exists between the CSI report time and the time when BS uses the CSI for traffic transmission. This temporal delay can result in that the reported CSI becomes outdated CSI, particular for mobile UEs, where the reported CSI no longer reflects actual channel conditions the UE experiences. The outdated CSI could degrade link performance and scheduling efficiency.

To address the challenges, 3GPP discusses the use of AI/ML on UE side for temporal CSI prediction. The AI/ML-based temporal CSI prediction aims to predict CSI for channel conditions associated with future time instances based on historic CSI measurements.

Studies were conducted in Release 18 and parts of Release 19 and the performance improvement is observed. For example, it was observed that user perceived throughput

(UPT) could improve by approximately 5% compared to non-AI/ML-based CSI prediction method that is introduced in Release 18 to improve performance loss for a UE at high / medium especially in MU-MIMO scenarios.

As a result, standardization of CSI prediction using UE-side model is proceeding in Release 19 starting from Q1 of 2025. This effort focuses on developing the functionality-based LCM procedures necessary to support data collection for AI/ML training, inference configuration and reporting, and performance monitoring. Additionally, the contents of the predicted CSI to be reported to BS, such as whether to reuse the Release 18 Type II Doppler codebook, are to be specified. For performance monitoring, discussions are ongoing regarding whether to introduce intermediate-KPI-based performance monitoring mechanism with squared generalized cosine similarity (SGCS) being considered as potential intermediate KPIs.

CSI reporting overhead is a challenge when the number of antennas and frequency resources is increased. To address the overhead, 3GPP also discusses the compression of CSI using AI/ML. Specifically, the UE compresses the CSI in the spatial and frequency domains using an AI/ML model, and reports the compressed information to the BS as CSI report. The BS then uses an AI/ML model on the network-side to reconstruct the original CSI from the compressed information, reported by the UE. In the studies conducted in Release 18, a reduction in CSI overhead of approximately 10% to 60% was observed compared to traditional CSI reporting methods, i.e., not using AI/ML [10].

CSI compression involves a two-sided model where inference processing using AI/ML models is executed on both UE side and network side. One of the challenges in considering two-sided model is how to coordinate training between UE side and network side. With this context, in Release 18, several types of training that involve different degrees of collaboration between UE side and network side were studied in terms of inter-vendor training collaboration complexity, performance, maintainability, and standardization impact. CSI compression using two-sided AI/ML model continues to be studied in Release 19 in order to alleviate / resolve the issue related to inter-vendor training collaboration. The issue on improving the trade-off between performance, computational complexity / overhead is also under discussion.

I-1.1.3. Beam Management

Especially in high-frequency bands, such as millimeter wave, beamforming operation is essential to extend coverage and maintain robust connectivity in a cell. Since Release 15, beam management (BM) has been supported in 3GPP new radio (NR) specifications. To extend the coverage of a beam while still covering the cell area, BS needs to apply a larger number of narrow beams. However, this increases the overhead of CSI reference signal (CSI-RS) transmission for BM. In addition, the inherent delay between beam reporting and beam utilization could lead to the use of outdated beam information, particularly for medium / high mobility UEs, resulting in inappropriate beam choices for traffic transmission.

To address these challenging BM issues, 3GPP has initiated discussions on applying AI/ML to BM. Two BM cases, 1) BM Case 1, i.e., “spatial-domain downlink beam prediction” and 2) BM Case 2, i.e., “temporal downlink beam prediction”, were studied and evaluated in Release 18.

For BM Case 1, optimal beam(s) are predicted from a set of downlink beams (“set A beams” based on measurement results from a set of downlink beams (“set B beams”) which is a subset of the “set A beams” or beams different from “set A beams”, This approach aims to reduce the overhead associated with massive CSI-RS transmission for beam measurement. For BM Case 2, optimal beam prediction is performed based on historic measurement results. This approach allows the AI/ML to capture and learn the evolution of channel conditions over time, thereby predicting optimal beams for future time instances. As evaluated during the study phase in Release 18, AI/ML can provide good beam prediction performance. For example, for BM Case 1, most evaluation results showed that 70% ~ 90% or even more than 90% beam prediction accuracy could be achieved by measuring only 1/4 of the beams, compared to measure all beams.

AI/ML for BM Case 1 and Case 2 can be implemented on either UE side or BS side. Standardization efforts are underway to specify the necessary signaling and procedure to support AI/ML training, inference, and performance monitoring. Existing CSI framework is reused to integrate the AI/ML for BM in the current 3GPP specification, ensuring minimal specification impact.

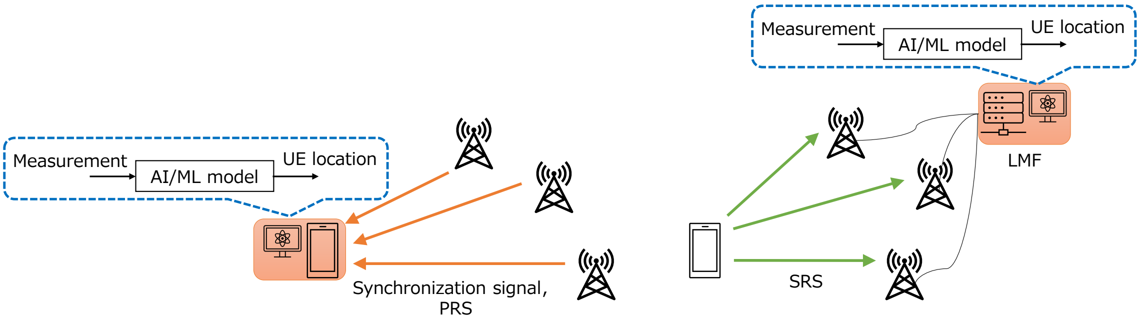
I-1.1.4. Positioning

In 3GPP, various positioning mechanisms has been specified for both downlink and uplink, including positioning using reference signals, positioning based on timing differences, positioning based on signal power, and positioning based on the angle of arrival of received signals, etc.

A key challenge in positioning is that the accuracy of location estimation heavily depends on whether measurements can be performed in line-of-sight (LOS) environments. In non-line-of-sight (NLOS) environment, such as indoor factories, or in environments with a high degree of multipath, the accuracy of positioning degrades.

Therefore, positioning accuracy improvement using AI/ML was considered. Specifically, two sub use cases were discussed: 1) location information is directly estimated using AI/ML models, and 2) AI/ML models generate intermediate statistical information for positioning estimation.

For direct location information estimation, two approaches were identified as illustrated in Fig. I-1.1.4-1: 1) performing training and inference on UE side, and 2) having the BS assist in training and inference through the location management function (LMF). Here, the LMF is a network function responsible for location information services as located in the 5G core network. In the studies in Release 18, it was found that the accuracy can be improved from 15 meters to below 1 meter using AI/ML models in indoor factory scenarios.



(a) UE-side model (b) BS-assisted LMF-side model

Fig. I-1.1.4-1: Application of AI/ML for positioning accuracy improvement

For intermediate statistical information for positioning estimation, to use timing and LOS / NLOS determination were discussed. It was studied whether training and inference would be performed solely on BS side or solely on UE side. In Release 18 study, a significant accuracy improvement was observed, comparable to the case of directly estimating location information.

Based on the Release 18 study, the positioning accuracy improvement is specified in Release 19. For direct estimation of location information using AI/ML models, both scenarios where training and inference are performed on UE side and where BS assists the training and inference through LMF, are to be specified. For generating intermediate statistical information using AI/ML models, scenarios where training and inference are conducted on BS side, are to be specified. Additionally, as the intermediate statistical information, at least information related to LOS / NLOS conditions and timing information, are to be used. Furthermore, signaling and mechanisms for LCM of AI/ML models are progressing. Additionally, methods for aligning the network-side additional conditions between the training and inference for UE-side inference are also under the discussion.

I-1.1.5. Mobility

To further enhance the mobility functions, AI/ML for mobility is currently under investigation as a study item in Release 19. The primary role of the mobility function is to manage the transition from the serving cell to a cell / gNB with higher quality based on the measurement results obtained from the UE, referred to as handover. AI/ML is expected to improve the efficiency of the mobility function, with two goals identified for this study, as illustrated in Fig. I-1.1.5-1 and I-1.1.5-2.

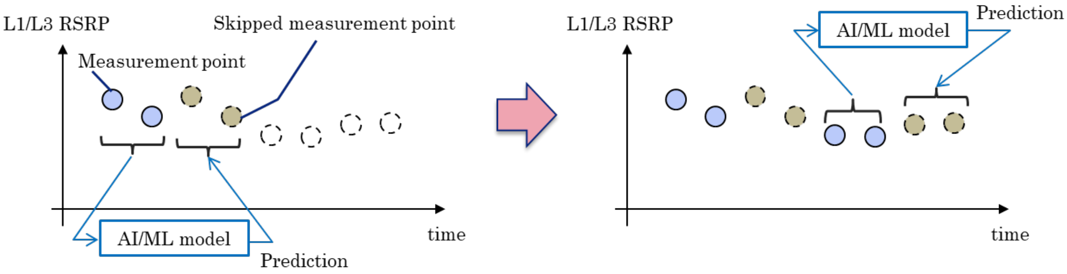


Fig. I-1.1.5-1: Measurement reduction for mobility

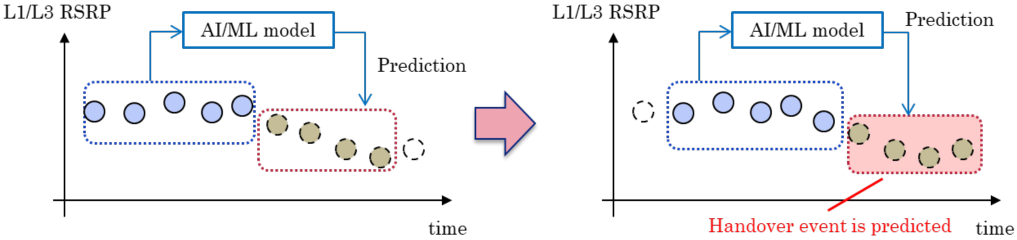
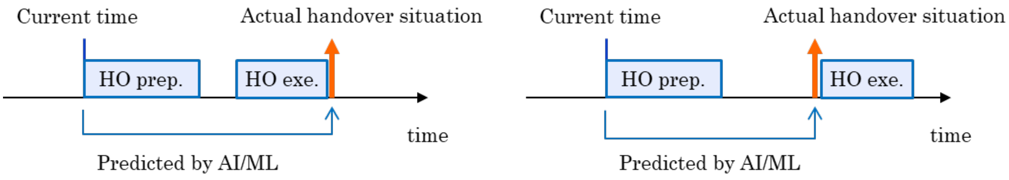


Fig. I-1.1.5-2: Handover event prediction

The first study goal is the reduction of measurement effort, shown in Fig. I-1.1.5-1. In this use case, UE skips measurements on certain resource that are usually measured for mobility functions. Instead of performing actual measurements, AI/ML predicts the received power at these resources and complements the measurement results accordingly. Consequently, UE can reduce the measurement effort if the accuracy of the prediction is sufficiently high. While Fig. I-1.1.5-1 outlines time domain predictions, investigations into predictions in the frequency and spatial domains are also underway.

The second study goal focuses on improving handover performance by prediction of handover situation. In Fig. I-1.1.5-2, AI/ML predicts whether a handover-related event is likely to occur in the future. Using the results of these predictions, a prediction-based handover process is considered, as depicted in Fig. I-1.1.5-3.



(a) Early handover execution (b) Early handover preparation

Fig. I-1.1.5-3: Prediction-based early handover

If a handover is executed preemptively based on predictions, as shown in Fig. I-1.1.5-3(a), the UE can switch to a neighboring cell before the quality of the current serving cell declines. This approach also mitigates the risk of handover failure due to sudden degradation of the current cell's quality. Conversely, since the execution of the handover depends on the prediction results, prediction errors could lead to improper handovers. To address these potential drawbacks, the use case illustrated in Fig. I-1.1.5-3(b) is also being considered. Handover preparation is initiated immediately after the prediction, but handover execution, such as handover command transmission, occurs only after the actual measurement captures the handover situation. Even in this case, the handover is completed earlier than in the legacy handover because the preparation is completed in advance.

In 3GPP Release 19, the effectiveness of the above prediction capabilities is evaluated and analyzed through simulations. Based on the study results, 3GPP will then identify the specification impact of AI/ML for mobility, with detailed specifications to be discussed in Release 20.

I-1.2. Standardization in O-RAN

The O-RAN Alliance aims to transform the way RAN is built by promoting openness, intelligence, and flexibility. Its mission is to drive the mobility industry towards an ecosystem of innovative, multi-vendor, interoperable, and autonomous RAN, with reduced cost, improved performance and greater agility. The Alliance has established technical working groups (WGs) focused on specific areas such as use cases, architecture, RAN intelligent controller (RIC), open fronthaul, cloudification, and security. The O-RAN architecture and interface specifications are consistent with 3GPP architecture and interface specifications to the extent possible.

In the O-RAN architecture [14], service management and orchestration (SMO) framework contains non-real-time RIC (Non-RT RIC) function which supports intelligent RAN optimization in non-real-time (i.e., greater than one second) by providing policy-based guidance. Non-RT RIC can leverage SMO services such as data collection and provisioning services of O-RAN nodes. Near real-time RIC (Near-RT RIC), O-CU-CP, O-CU-UP, O-DU, and O-RU are the network functions for the radio access side. Near-RT RIC enables control and optimization of O-RAN (O-CU and O-DU) nodes and resources with near real-time control loops (i.e., 10 ms to 1 s), The Near-RT RIC collects near real-time RAN information from the O-RAN nodes and controls the behaviors of them on the basis of the policies and the enrichment data provided by the Non-RT RIC.

Potential O-RAN use cases are discussed in O-RAN WG1 use case task group (UCTG) [15]. The use cases are described at a high level, emphasizing how the use case is enabled by O-RAN architecture. These high-level use cases are prioritized within O-RAN, and selected use cases are further detailed in O-RAN WG1 UCTG and relevant O-RAN WGs to define the requirements for O-RAN components and their interfaces.

One of the key innovations driven by O-RAN is the concept of intelligent RAN. By integrating AI/ML into the network, operators can improve performance, optimize resource allocation, and enhance user experiences. AI/ML workflow technical report (TR) was created within WG2, summarizing the deployment scenarios, procedures, requirements, and issues for AI/ML [16]. Based on the requirements outlined in this TR, "AI/ML in O-RAN" was established as a feature of MVP-C (Minimum Viable Plan Committee) to specify the architecture and interfaces necessary to realize the AI/ML lifecycle using RIC. In the following sections, we provide overview of the AI/ML framework in O-RAN and, several O-RAN use cases that utilize AI/ML in [15].

I-1.2.1. AI/ML Framework

This section provides the framework of AI/ML procedure in O-RAN [16]. The potential mapping relationship between the ML components and network functions, interfaces defined in O-RAN are illustrated in Fig. I-1.2.1-1.

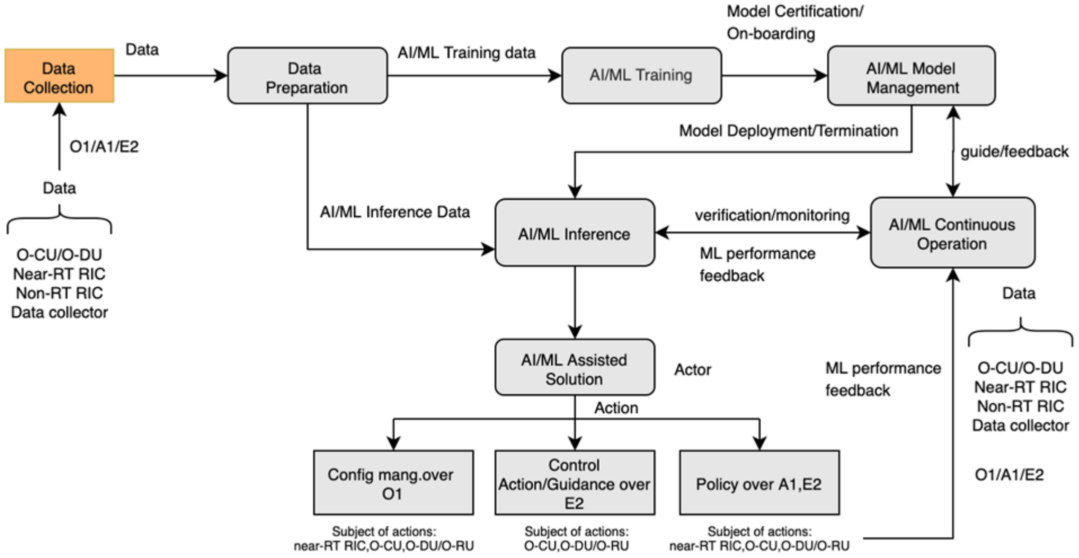


Fig. I-1.2.1-1: AI/ML framework

The Non-RT RIC and Near-RT RIC support AI/ML workflow services. The following AI/ML services have been defined:

- AI/ML training services: These services allow an AI/ML training service Consumer to request training of an AI/ML model by specifying training requirements (e.g., required data, model, validation criteria, etc.).

- AI/ML model management and exposure services: These services enable
  - AI/ML model registration / deregistration
  - AI/ML model discovery
  - AI/ML model change subscription
  - AI/ML model storage
  - AI/ML model training capability registration / deregistration (optional service)
  - AI/ML model training capability query (optional service)
  - AI/ML model retrieve
- AI/ML model performance monitoring services: These services allow an authorized AI/ML performance monitoring service Consumer to request monitoring the performance of a deployed AI/ML model. The performance information of an AI/ML model is produced by an App within which the model is deployed or by AI/ML model inference service Producer performing the model inference.
- AI/ML model inference services: These services allow an App to request and or to cancel the inference for a registered AI/ML model. The App needs to be authorized to request inference for registered AI/ML models.

The ML functions are implementation variability components, there are many combinations of the deployment scenarios. The typical deployment scenarios that are considered for AI/ML framework in O-RAN are:

- Deployment Scenario 1.1: AI/ML Continuous Operation / AI/ML Model Management / Data Preparation / AI/ML Training and AI/ML Inference are all in Non-RT RIC.
- Deployment Scenario 1.2: AI/ML Continuous Operation / Data Preparation (for training) / AI/ML Training are in Non-RT RIC, AI/ML Model Management is out of Non-RT RIC (in or out of SMO). Data Collection (for inference) / Data Preparation (for inference) / AI/ML Inference is Near-RT RIC.
- Deployment Scenario 1.3: AI/ML Continuous Operation / AI/ML Inference are in Non-RT RIC. Data Preparation / AI/ML Training / AI/ML Model Management are out of Non-RT RIC (in or out of SMO).
- Deployment Scenario 1.4: Non-RT RIC acts as the ML training host for offline model training and the Near-RT RIC as the ML training host for online learning and ML inference host.

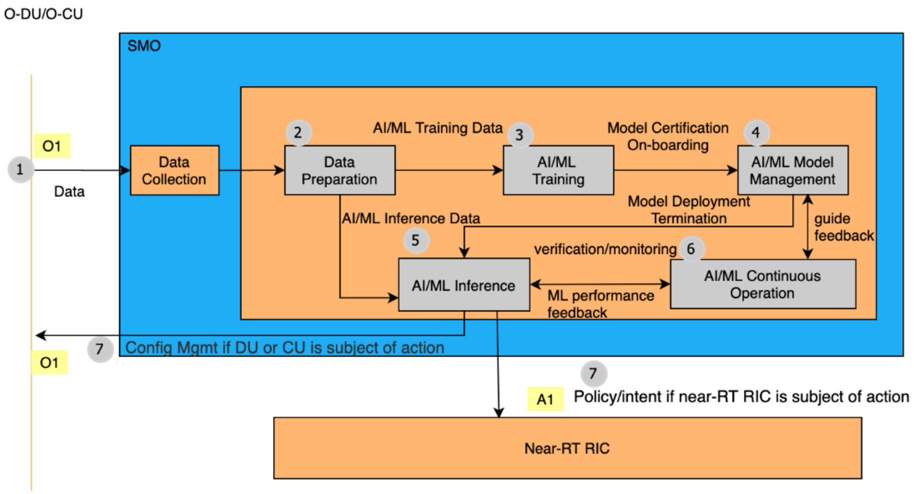


Fig. I-1.2.1-2: Deployment Scenario 1.1

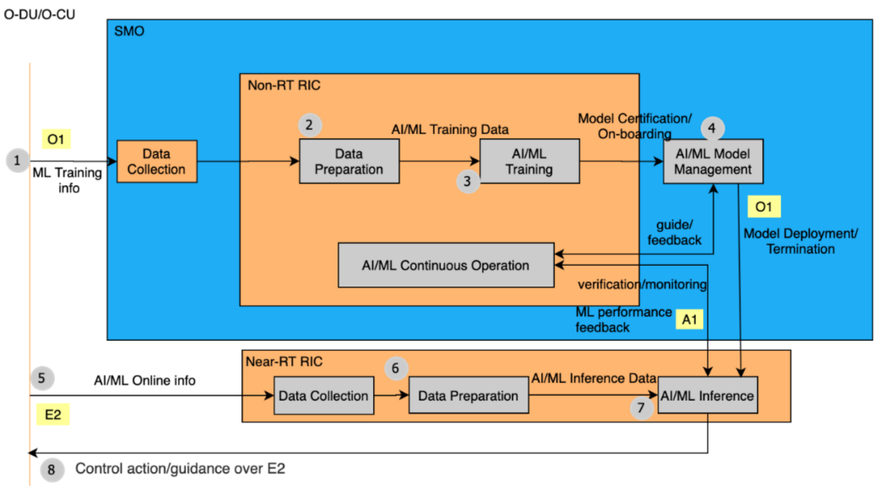


Fig. I-1.2.1-2: Deployment Scenario 1.2

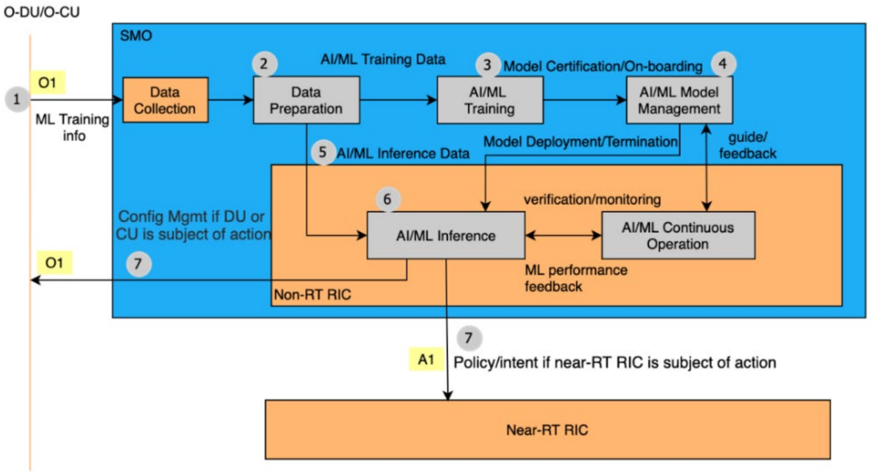


Fig. I-1.2.2: Deployment Scenario 1.3

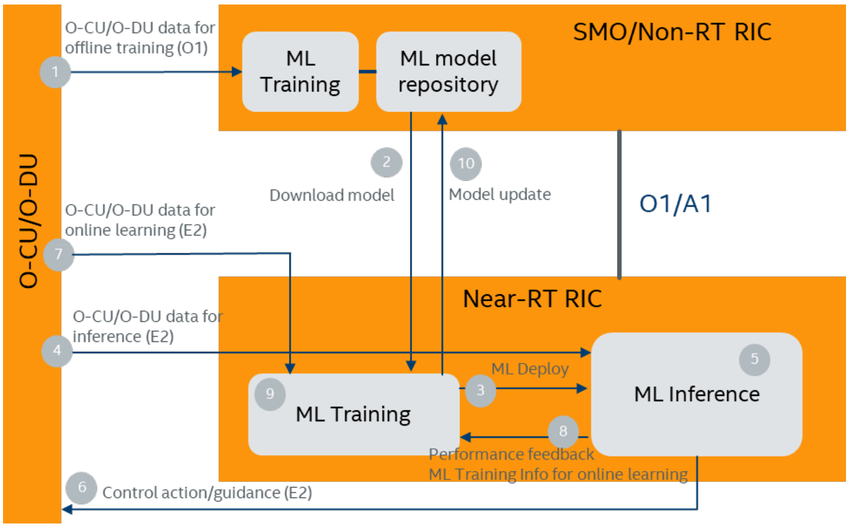

O-CU/O-DU
O-CU/O-DU data for offline training (O1)
1
ML Training
ML model repository
SMO/Non-RT RIC
2
Download model
10
Model update
O1/A1
O-CU/O-DU data for online learning (E2)
7
Near-RT RIC
O-CU/O-DU data for inference (E2)
4
5
ML Inference
ML Deploy
3
9
ML Training
8
Performance feedback
ML Training Info for online learning
6
Control action/guidance (E2)

Fig. I-1.2.1-5: Deployment Scenario 1.4

I-1.2.2. Massive MIMO Beamforming Optimization

Massive MIMO (mMIMO) is a crucial technology for 5G, leveraging multi-antenna transmission and reception to improve power levels and enhance capacity by spatial multiplexing operations. In addition, advantages include advanced network management technologies like beam shaping, beam-based load balancing, optimized beam mobility, adaptive cell coverage areas. In order to optimize networks, fully digital beamforming (BF) methods are to be employed for below 6 GHz frequency. Grid of Beams (GoB) is a BF method which aims at selectively covering regions of interest with a suitable subset of radio beams. Beam-based mobility robustness optimization is a BF method enhancing beam specific mobility performance, e.g., by adding beam specific individual offsets.

The high number of configuration parameters, the amount of measurement input data, the complexity, pro-activeness as well as non- and near-real time requirements suggest the application of AI/ML techniques. In this use case, three optimization loops for mMIMO BF were proposed.

1) Non-RT massive MIMO GoB beamforming optimization

   The concept of Non-RT BF optimization is shown in Fig. I-1.2.2-1. Non-RT RIC hosts an application with long-term analytics function (= ML training), whose task is to collect, process and analyze antenna array parameters, cell performance KPIs, UE mobility / spatial density data, traffic density data, interference data and BF gain / beam reference signal received power (RSRP) and minimization of drive tests (MDT) measurement data. The output of the BF optimization inference can be optimized BF configuration, number of beams, beam elevation, beam horizontal & vertical widths and power allocation of beams.

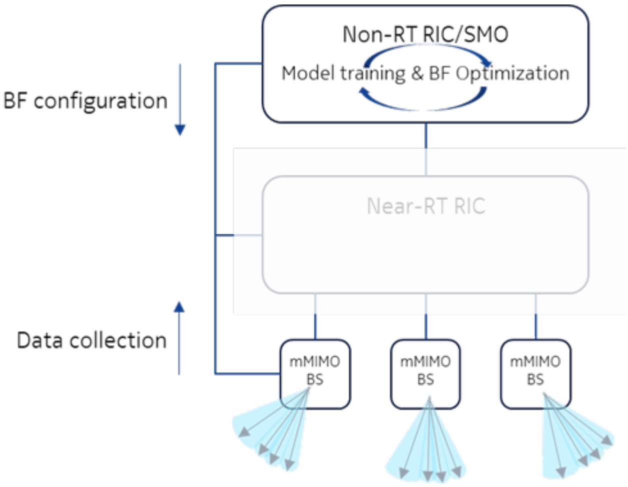


Fig. I-1.2.2-1: Non-RT BF optimization

2) Near-RT massive MIMO beam-based Mobility Robustness Optimization (bMRO)
The concept of bMRO is shown in Fig. I-1.2.2-2. Non-RT RIC hosts an application with long-term analytics function (= ML training), whose task is to collect and analyze underlying GoB configuration, if GoB configuration exists, beam mobility and failure statistics, L1 / L2 RSRP values, potential source-target beam pairs. Near-RT RIC hosts an xApp with bMRO optimization function (= ML inference), whose task is to monitor potential source-target beam pairs and optimize beam mobility for scheduling by managing user-beam paring. The output of the bMRO optimization function can be adjusted offsets for candidate source-target beam pairs for beam mobility.

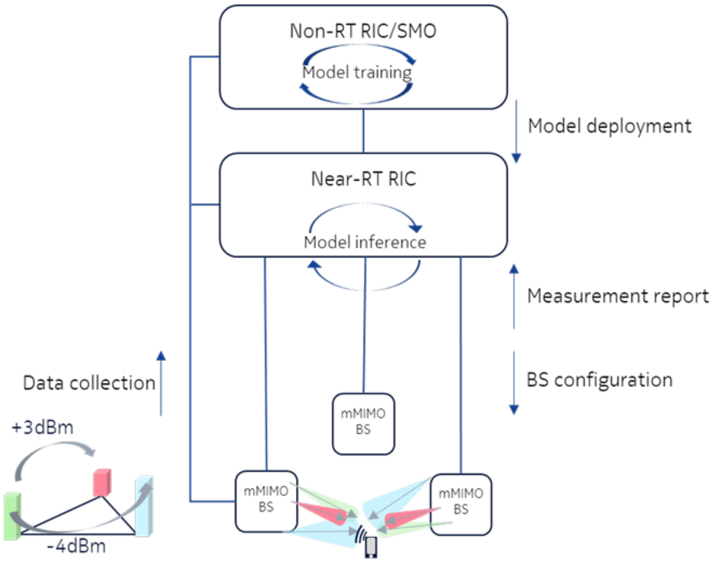


Fig. I-1.2.2-2: Near-RT beam-based mobility robustness optimization

3) Near-RT massive MIMO Beam Selection Optimization (BSO)

The concept of BSO function is shown in Fig. I-1.2.2-3. Non-RT RIC hosts an application with long-term analytics function (= ML training), whose task is to collect and analyze underlying GoB configuration, if GoB configuration exists, beam mobility and failure statistics, L1 / L2 RSRP values, potential source-target pairs. Near-RT RIC hosts an xApp with BSO function (= ML inference), whose task is to monitor potential source-target beam pairs, and to optimize beam mobility for scheduling by managing user-beam pairing. The output of the BSO optimization function can be adjusted offsets for candidate source-target beam pairs for beam mobility.

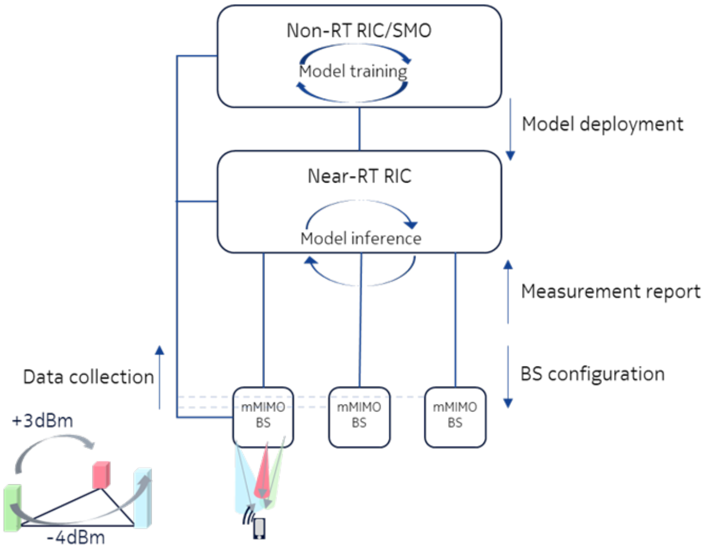


Fig. I-1.2.2-3: Near-RT BSO function

I-1.2.3. RAN Slice SLA Assurance

The 3GPP standards architected a sliceable 5G infrastructure which allows creation and management of customized networks to meet specific service requirements that can be demanded by future applications, services and business verticals. Such a flexible architecture needs different requirements to be specified in terms of functionality, performance and group of users which can greatly vary from one service to the other. The 5G standardization efforts have gone into defining specific slices and their Service Level Agreements (SLAs) based on application / service type. Since network slicing is conceived to be an end-to-end feature that includes the core network, the transport network and the RAN, these requirements should be met at any slice subnet [17].

The requirements of network slicing in RAN include customizable network capabilities such as the support of very high data rates, traffic densities, service availability and very low latency. These capabilities are always provided based on an SLA between the mobile operator and the business customer, which brought up interest for mechanisms to ensure slice SLAs and prevent its possible violations. O-RAN's open interfaces and AI/ML-based architecture will enable such challenging mechanisms to be implemented and realize the network slicing in an efficient manner. This use case was proposed to clarify necessary mechanisms and parameters for RAN slice SLA assurance.

As shown in Fig. I-1.2.3-1, RAN slice SLA assurance scenario involves Non-RT RIC, Near-RT RIC, E2 nodes and SMO interaction. The scenario starts with the retrieval of RAN specific slice SLA / requirements (possibly within SMO or from NSSMF depending on operator deployment options). Based on slice specific performance measurements from E2 nodes, Non-RT RIC and Near-RT RIC fine-tune RAN behavior aligned with O-RAN architectural roles to assure RAN slice SLAs dynamically. Non-RT RIC monitors long-term trends and patterns for RAN slice subnets' performance and employs AI/ML methods to perform corrective actions through SMO (e.g., reconfiguration via O1) or via creation of A1 policies. Non-RT RIC can also construct / train relevant AI/ML models that will be deployed at Near-RT RIC. A1 policies possibly include scope identifiers (e.g., S-NSSAI) and statements such as KPI targets. On the other hand, Near-RT RIC enables optimized RAN actions through execution of deployed AI/ML models in near real-time by considering both O1 configuration (e.g., static RRM policies) and received A1 policies, as well as received slice specific E2 measurements.

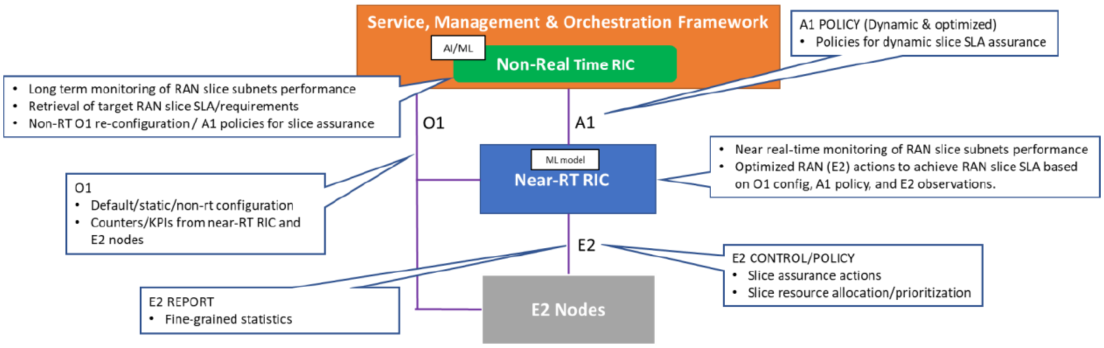

Service, Management & Orchestration Framework
AI/ML
Non-Real Time RIC
A1 POLICY (Dynamic & optimized)
• Policies for dynamic slice SLA assurance
• Long term monitoring of RAN slice subnets performance
• Retrieval of target RAN slice SLA/requirements
• Non-RT O1 re-configuration / A1 policies for slice assurance
O1
A1
ML model
Near-RT RIC
• Near real-time monitoring of RAN slice subnets performance
• Optimized RAN (E2) actions to achieve RAN slice SLA based on O1 config, A1 policy, and E2 observations.
O1
• Default/static/non-rt configuration
• Counters/KPIs from near-RT RIC and E2 nodes
E2
E2 CONTROL/POLICY
• Slice assurance actions
• Slice resource allocation/prioritization
E2 REPORT
• Fine-grained statistics
E2 Nodes

Fig. I-1.2.3-1: Slice SLA assurance

I-1.2.4. Energy Saving

Energy saving (ES) of the RAN is an important topic for network operators. ES for legacy and 5G networks can be carried out using manual configuration in different network layer and in different time scales. However, due to the varying nature of traffic load and to user mobility, the optimization of energy consumption of the RAN is complex. There is a risk that RAN equipment consume much energy while serving low traffic, or even no traffic at all.

O-RU is responsible for the major part of energy consumption in the mobile network. 3GPP defines both centralized and distributed ES features [18], which are mainly targeting intra- or inter-RAT cell on/off switching, The ES use case was proposed to leverage on O-RAN AI/ML services and open interfaces in order to introduce optimized ES solutions involving switching off/on of different network components at different time scale. The ES use cases is divided into three sub use cases.

1) Carrier and cell switch off/on ES

   Time scale: non-real-time for both control and controlled system. The feature aims at reducing O-CU / DU / RU power consumption by switching off/on one or more carriers or a cell of a given technology. AI/ML assisted solutions in the Non-RT RIC can be used to control the traffic load of the carriers and the cell, and to automatically decide when to switch off/on one or more carriers or a cell using O1 and/or open fronthaul M-plane parameter configurations. Off/on switching is accompanied with adequate traffic steering, guided by policies, to ensure service continuity and quality of service.

2) RF channel switch off/on ES

   Time scale: non- or near real-time are possible for both control and controlled system. This feature aims at reducing power consumption of O-RU with massive MIMO deployment by switching off/on certain RF channels. Using AI/ML assisted solutions, rApp or xApp will trigger switching off/on certain RF channels, based on traffic information such as load, user location and mobility. As example, one can switch off 32 out of 64 RF channels in a digital mMIMO architecture or reduce the number of layers and/or number of multi-user scheduled UEs in a hybrid architecture. The O-RU reconfiguration can be performed using the open fronthaul M-plane from E2 node or SMO.

3) Advanced sleep mode ES

   Time scale for control: near real-time. Time scale for the controlled system: real-time and near real-time. This feature is expected to reduce power consumption by partially switching off O-RU components. Using multi-dimensional data, e.g., traffic load, user service type, energy efficiency measurements, etc., the Near-RT RIC can configure cell parameters, such as the SSB periodicity needed for the operation of advanced sleep modes.

REFERENCE

[1] X. Lin, L. Kundu, C. Dick, and S. Vekayutham, “Embracing AI in 5G-Adcaned Toward 6G: A Joint 3GPP and O-RAN Perspective,” IEEE Communications Standards Magazine, Vol.7, No.4, p.76 – 83, December 2023.

[2] X. Lin, “Artificial Intelligence in 3GPP 5G-Advanded: A Survey,” IEEE Communication Society, ComSoc Technology News (CTN), September 2023.

[3] W. Chen, X. Lin, J. Lee, A. Toskala, S. Sun, C. F. Chiasserini, and L. Liu, “5G-Advanced Toward 6G: Past, Present, and Future,” IEEE Journal of Selected Areas in Communications, Vol.41, No.6, p.1592 – 1619, June 2023.

[4] 3GPP TR 37.817, “Study on enhancement for Data Collection for NR and EN-DC,” V17.0.0, April 2022.

[5] RP-233441, “Revised WID: Artificial Intelligence (AI) / Machine Learning (ML) for NG-RAN,” CMCC, Ericsson, December 2023.

[6] RP-233442, “WI summary for WI Artificial Intelligence (AI) / Machine Learning (ML) for NG-RAN,” CMCC, Ericsson, December 2023.

[7] RP-234054, “New SID: Study on enhancements for Artificial Intelligence (AI) / Machine Learning (ML) for NG-RAN,” China Unicom, December 2023.

[8] RP-250812, “Summary for RAN Release 20 5G-Adv,” March 2025.

[9] RP-221348, “Revised SID: Study on Artificial Intelligence (AI) / Machine Learning (ML) for NR Air Interface,” Qualcomm, June 2022.

[10] 3GPP TR 38.843, “Study on Artificial Intelligence (AI) / Machine Learning (ML) for NR air interface,” V2.0.1, December 2023.

[11] RP-243244, “Revised WID: Artificial Intelligence (AI) / Machine Learning (ML) for NR Air Interface,” Qualcomm, December 2024.

[12] RP-243245, “New SID: Study on Artificial Intelligence (AI) / Machine Learning (ML) for NR air interface Phase 2,” Qualcomm, December 2024.

[13] RP-234055, “New SID: Study on Artificial Intelligence (AI) / Machine Learning (ML) for mobility in NR,” CMCC, December 2023.

[14] O-RAN Alliance, “O-RAN architecture description,” O-RAN Alliance, Tech. Rep. O-RAN.WG1.O-RAN-Architecture-Descriptionv07.00, 2022.

[15] O-RAN Alliance, “Use Cases Analysis Report,” O-RAN Alliance, Tech. Rep. O-RAN.WG1.TR.Use-Cases-Analysis-Report-R004-v15.00, 2024.

[16] O-RAN Alliance, “AI/ML workflow description and requirements,” O-RAN Alliance, Tech. Rep. AI/ML workflow description and requirements, 2021.

[17] 3GPP TS 23.501: “System Architecture for the 5G System (5GS); Stage 2”, Release 16, December 2019.

[18] 3GPP TS 28.310: “Management and orchestration; Energy efficiency of 5G”, Release 17, December 2021.

I-2. Introduction of AI and Digital Twin Technologies for 6G

Tetsuya Yamamoto, Panasonic Holdings Corporation
Takahiro Yamazaki, NTT Network Innovation Laboratories, NTT Corporation

I-2.1. AI for Signal Processing / Air-interface

In recent years, with the increasing complexity of next-generation wireless communication systems such as Beyond 5G/6G, environments with numerous interdependent parameters that are difficult to manage with conventional methods have become a reality. This is driven by the challenge of achieving high-precision and real-time performance across a variety of communication functions, including channel estimation, beamforming, positioning, and resource allocation in time, frequency, space, and power, etc. In response, AI/ML technologies are expected to provide groundbreaking solutions by solving complex nonlinear mapping problems and analyzing vast amount of data.

In wireless signal processing at the physical layer, efforts are underway to replace conventional processes such as channel coding, synchronization, channel estimation, beamforming, and transmit power control with AI/ML models like deep neural networks. For example, AI/ML-based optimization are suggested to contribute to reduce computational complexity and improved accuracy in signal detection, blind channel estimation and demodulation using minimal reference signals, and the decoding of advanced error-correcting codes. As a result, optimal signal processing is expected to be maintained even in the face of environmental variations, noise, and interference [1 – 3].

As shown in Section I-1.1, there are ongoing standardization efforts to exploit AI/ML in the air interface. The 3GPP has been studying the application of AI/ML to NR air interface since Release 18, In Release 19, the specification of CSI prediction, beam management, and positioning are being specified, and the feasibility of CSI compression is being studied.

Furthermore, the concept of an AI-native air interface represents an evolution from traditional, fixed air interface protocols toward new communication methods that dynamically adapt to the constraints and variability of the wireless environment as well as to hardware imperfections. As a first step, a hybrid system combining AI/ML and non-AI/ML processing blocks such as signal detection, channel estimation, and symbol mapping at the physical layer is expected. Looking ahead, it is conceivable that AI/ML models integrating multiple functions, such as joint channel estimation, equalization, and de-mapping, will emerge. With advancements in hardware acceleration and improvements in the reliability of AI/ML models themselves, research is moving towards the realization of system composed entirely of AI/ML components.

Moreover, in high-frequency ranges such as millimeter-wave and terahertz bands, RF impairments, such as the nonlinearity of power amplifiers (PAs), frequency selectivity, IQ imbalance, direct current (DC) offset, carrier leakage, and phase noise, have a significant impact on system performance. For the nonlinear distortion of PA, nonlinear compensation techniques, such as digital predistortion employing AI/ML, have attracted considerable attention. By utilizing neural networks, high compensation effects are anticipated even for complex nonlinear distortions that conventional polynomial-based models cannot adequately express, although new challenges in terms of computational resources and hardware implementations have also surfaced. Technologies that compensate for multiple RF impairments by utilizing AI/ML such as deep neural networks have also attracted attention.

These advances in AI/ML technologies are exerting a profound influence on the overall design of wireless communication systems, leading to the establishment of performance requirements such as training / inference accuracy and latency KPIs from both communication and AI/ML perspective. in 6G networks, in order to meet these KPIs, support for large-scale distributed learning and real-time inference will be essential, along with the integrated system design that transcends traditional boundaries between communication and AI.

In summary, AI/ML is set to revolutionize conventional signal processing and air interface paradigms, serving as the key technology to achieve dynamic and highly efficient optimization in complex wireless environments. It is poised to become a central component of future Beyond 5G/6G systems.

I-2.2. AI for RAN

AI/ML technologies are expected to be used for operations, administration and maintenance (OAM) and dynamic control of RAN.

For OAM of RAN, instead of a manual parameter configuration, an automatic parameter configuration by AI/ML technologies is proposed by [4, 5]. It will reduce human operation resources and human errors.

For dynamic control of RAN, dynamic traffic offloading, resource allocation and power control by AI/ML technologies are proposed by [6, 7]. It will improve quality of communication and power efficiency.

On the other hand, as described in previous section (I-1-2), the architecture of AI/ML for RAN is standardized in O-RAN Alliance [8, 9]. With O-RAN RIC, application-aware RAN control such as application-based resource allocation will be enabled.

Additionally, AI/ML technologies are used for system failure detection [10]. With AI/ML, the threshold for failure detection can be dynamically configured, and it makes the probability of failure detection higher.

In summary, AI/ML technologies are expected to reduce operation costs of RAN, to improve quality of communication of RAN. Following 5G system, it will also become important for Beyond 5G and 6G systems.

I-2.3. AI for Radio Propagation / Digital Twin

AI/ML technologies are expected to evolve radio propagation and radio simulation in digital virtual environments such as digital twin.

For radio propagation, as described in [11, 12], AI/ML technologies can be applied for channel parameter estimation, channel modelling, channel prediction and LOS / NLOS identification. AI/ML technologies can make ML models with multimodal values such as measured received signal strength indicator (RSSI), geographical information, camera images, states of UEs and etc. Using this ML model, it is expected to support more flexible radio propagation situations and scenarios for Beyond 5G and 6G.

For radio simulation in digital twin, AI/ML technologies can reduce computational cost while maintaining simulation accuracies. In [13], for real-time digital twin system, a ML model trained with ray tracing results, geographical data and rough propagation model is proposed. This paper shows that proposed model reduces computational cost while maintaining accuracies compared to conventional ray tracing. Not only this example, but also many AI/ML approaches have been investigated to implement more realistic and cost-effective radio simulation.

In summary, AI/ML is a promising approach to make close radio propagation model to the real one. It should accelerate the development of digital twin and cyber physical system.

I-2.4. Network Architecture for AI/ML Usage in RAN

Network architecture for using AI/ML for RAN is under consideration.

Conventionally, AI/ML application functions are placed in core network or cloud infrastructure. However, it increases the latency due to AI/ML processing is carried out in a location farther away than the cellular area provided by the RAN. To resolve this problem, placing AI/ML application functions to RAN side, such as MEC or computing infrastructure for vRAN, is proposed [14].

Additionally, for 6G, a network architecture which distributes AI/ML application functions to core network, RAN, user devices and all of the network functions is proposed. It will enable an adaptive computing resource allocation for AI/ML in end-to-end communications. With this architecture, more AI/ML applications will be effectively utilized in 6G systems.

REFERENCE

[1] 6G Flagship, University of Oulu, “6G visions on paper,” http://www.6gflagship.com/white-papers/

[2] J. Du, C. Jiang, J. Wang, Y. Ren, and M. Debbah, “Machine learning for 6G wireless networks: Carrying forward enhanced bandwidth, massive access, and ultrareliable/low-latency service,” IEEE Veh. Technol. Mag., Vol.15, No.4, pp.122-134, Dec. 2020.

[3] T. Ohtsuki, “Machine learning in 6G wireless communications,” IEICE Trans. Commun., Vol.E106-B, No.2, pp.75-83, Feb. 2023.

[4] Softbank Corp., “SoftBank Corp. Develops a Foundational Large Telecom Model (LTM),” Mar. 2025. https://www.softbank.jp/en/corp/news/press/sbkk/2025/20250319_03/

[5] KDDI Research Inc., “Successful Verification Experiment of Network Operation Using Dialogue with AI,” Feb. 2025. https://newsroom.kddi.com/english/news/detail/kddi_nr-467_3745.html

[6] Kyocera Corp., “Kyocera Develops AI-Powered 5G Virtualized Base Station For the Telecommunication Infrastructure Market,” Feb. 2025. https://global.kyocera.com/newsroom/news/2025/001001.html

[7] KDDI Corp., Nokia Solutions and Networks Japan G.K., “KDDI and Nokia have agreed to conduct a demonstration test that will reduce power consumption at base stations by up to 50% using AI control, the first of its kind in Japan,” Jun. 2021. https://news.kddi.com/kddi/corporate/newsrelease/2021/06/18/5193.html

[8] M. Nakajima, et. al., “Cognitive Foundation (CF) collaborative infrastructure technology that realizes the intelligent operation management of wireless access networks (RAN),” NTT Technical Journal, Nov. 2024. https://doi.org/10.60249/24115004

[9] T. Katsuragawa, et. al., “Initiatives toward Intelligent RAN,” NTT DOCOMO Technical Journal, Vol. 24, No.1, Mar. 2023. https://www.docomo.ne.jp/english/corporate/technology/rd/technical_journal/bn/vol24_1/003.html

[10] KDDI Corp., “Started the operation of a disability detection system utilizing AI,” Jan. 2024. https://newsroom.kddi.com/news/detail/kddi_pr-1097.html

[11] C. Huang et al., "Artificial Intelligence Enabled Radio Propagation for Communications—Part I: Channel Characterization and Antenna-Channel Optimization," in IEEE Transactions on Antennas and Propagation, vol. 70, no. 6, pp. 3939-3954, Jun. 2022. https://doi.org/10.1109/TAP.2022.3149663

[12] C. Huang et al., "Artificial Intelligence Enabled Radio Propagation for Communications—Part II: Scenario Identification and Channel Modeling," in IEEE Transactions on Antennas and Propagation, vol. 70, no. 6, pp. 3955-3969, Jun. 2022.
https://doi.org/10.1109/TAP.2022.3149665

[13] A. Saeizadeh et al., "AI-Assisted Agile Propagation Modeling for Real-Time Digital Twin Wireless Networks," 2024 IEEE 29th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Athens, Greece, 2024, pp. 1-6.
https://doi.org/10.1109/CAMAD62243.2024.10942746

[14] Softbank Corp., “SoftBank Corp. Announces Development of AITRAS, a Converged AI-RAN Solution,” Nov. 2024.
https://www.softbank.jp/en/corp/news/press/sbkk/2024/20241113_06/

II. Recent Activities of AI/ML and Digital Twin in Japan

This section introduces leading-edge R&D efforts on AI/ML and digital twin for Beyond 5G/6G in Japan, with the aim of accelerating R&D to advance future communications and services. This white paper on AI/ML and digital twin technologies includes 19 papers categorized into 4 typical technological categories, as shown below:

AI for signal processing / air interface

1. Scalable AI/ML for radio cellular access
2. Study on training collaboration at UE- / network-side for CSI compression with two-sided AI/ML model
3. Proof-of-concept for AI-native air interface toward 6G
4. Neural network-based digital pre-distortion for wideband power amplifier using DeepShift
5. AI calibration network under hardware limitations
6. Performance requirements and evaluation mythology for AI and communication in 6G

AI/ML for RAN

7. Study on AP clustering with deep reinforcement learning for cell-free massive MIMO
8. Cross-layer access control techniques using AI
9. AI-based application-aware RAN optimization
10. AIOps for autonomous networks
11. Logic-oriented generative AI technology for autonomous networks
12. In-network learning for distributed RAN AI ~Distributed LLMs via latent structure distillation~

AI/ML for radio propagation and digital twin

13. Throughput prediction technology for 28-GHz channels using physical space information
14. AI/ML-based radio propagation prediction technology
15. AI-based radio propagation modeling for wireless emulator
16. 6G simulator utilizing future prediction control technology based on AI/ML
17. Optimization of 6G radio access using digital twin
18. Digital-twin for and by Beyond 5G

Network architecture for AI/ML usage in RAN

19. Task-oriented 6G native-AI network architecture

II-1. 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.**

II-1.1. 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.

II-1.2. 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.

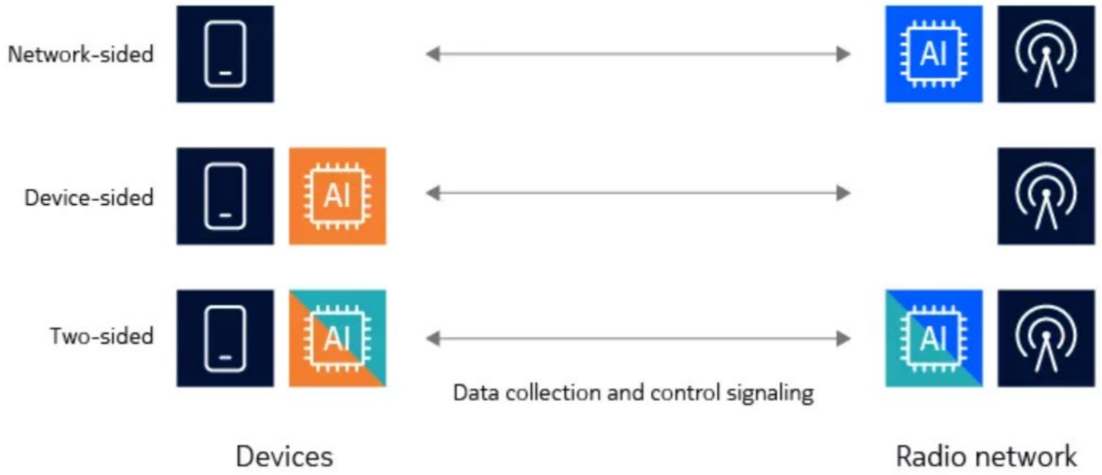


Fig. II-1.2-1: One-sided and two-sided AI/ML solutions in the mobile network air interface [3]

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

II-1.3. 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. II-1.3-1). 3GPP defines 114 values for *P0* and 8 values for α_{pl} resulting in 912 possible combinations.

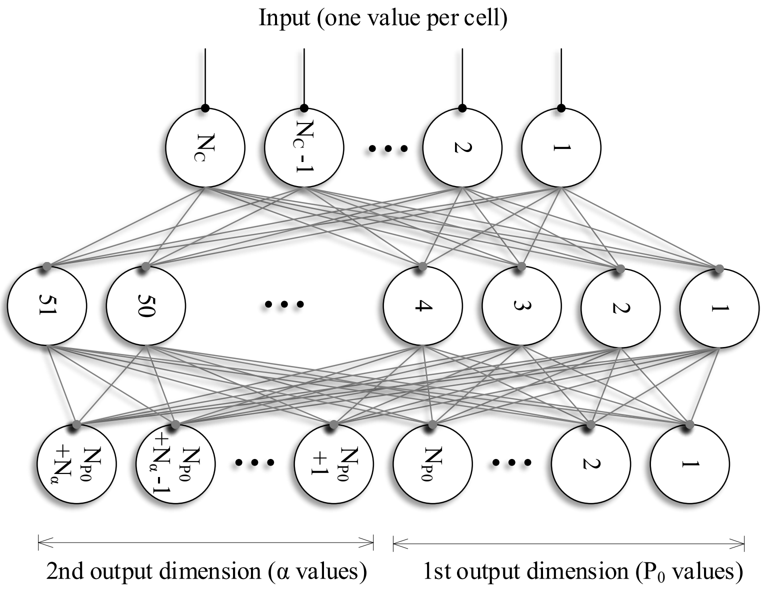

Input (one value per cell)
N_C
N_C-1
2
1
51
50
4
3
2
1
N_P0 +N_α
N_P0 +N_α-1
N_P0 +1
N_P0
2
1
2nd output dimension (α values)
1st output dimension (P_0 values)

Fig. II-1.3-1: 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. II-1.3-2) 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.

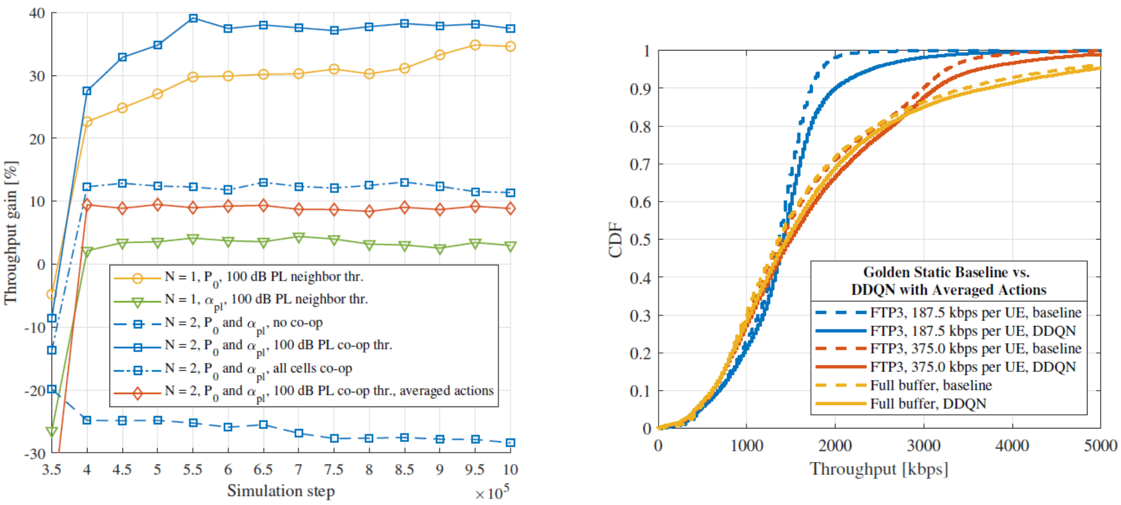


Fig. II-1.3-2: Simulation results: (left) Average Cell Throughout Gain, where N=2 refers to learning both *P0* and α_{PL} output dimensions; (right) Windowed User Throughput Distribution of with different offered loads [1]

II-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 LCM 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)

II-2. Study on Training Collaboration at UE- / Network-side for CSI Compression with Two-sided AI/ML Model

Tetsuya Yamamoto, Yasuaki Yuda, Hidetoshi Suzuki
Panasonic Holding Corporation
Maki Sugata, Tadashi Yoshida
Panasonic System Networks R&D Lab. Co., Ltd.

***Abstract*— In the 3rd Generation Partnership Project (3GPP), the application of artificial intelligence / machine learning (AI/ML) to the radio interface has been studied since Release 18. Channel State Information (CSI) compression is one of use cases studied in 3GPP. CSI compression involves a two-sided model where inference processing using AI/ML models is executed on both the user equipment (UE)-side and the network (NW)-side, which will be used as a starting point for studying two-sided AI/ML solutions in 6G. This paper provides an overview of training collaboration for CSI compression using a two-sided model and reports on the performance of several training collaboration approaches.**

II-2.1. Introduction

In the 3GPP, the application of AI/ML to the radio interface has been studied since Release 18 [1]. In Release 19, the actual standardization work is on the progress to realize use cases, such as beam management, positioning, and CSI prediction, where the introduction of AI/ML technologies would be effective, based on the study from Release 18 [2]. In addition, use case that were not concluded in Release 18 such as CSI compression and new use cases are also to be studied in Release 19 [3, 4].

CSI compression involves a two-sided model where inference processing using AI/ML models is executed on both the UE-side and the NW-side, and how to coordinate training between the UE-side and the NW-side is being studied.

In this paper, we provide the overview of the training collaboration for CSI compression using two-sided model and report on the performance of several training collaboration approaches being considered in 3GPP Release 19.

II-2.2. CSI Compression with Two-sided AI/ML Model

In downlink transmission, the base station (BS) needs to know the reception quality and propagation channel information of the UE to perform resource allocation and multiple-input multiple-output (MIMO) precoding. In the downlink transmission in new radio (NR) interface, the UE measures the reception quality and propagation channel state from the reference signal transmitted by the BS and reports the measurement results to the BS. This is the reporting of CSI from the UE.

CSI reporting overhead is a challenge when the number of antennas and frequency resources is increased. To address this overhead, the compression of CSI using AI/ML is being considered. Specifically, as shown in Fig. II-2.2-1, the UE compresses the CSI in the spatial and frequency domains using an AI/ML model, and reports the compressed information to the BS as CSI report. The BS then uses an AI/ML model on the NW-side to reconstruct the original CSI from the compressed information, reported by the UE. AI/ML models such as convolutional neural networks (CNN) and Transformers can be utilized for this purpose [5, 6]. In the studies conducted in Release 18, a reduction in CSI overhead of approximately 10% to 60% was observed compared to traditional CSI reporting methods i.e., not using AI/ML [7].

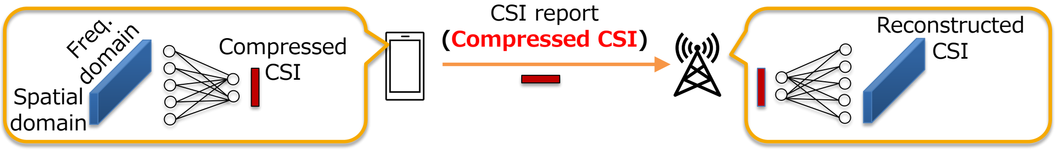


Fig. II-2.2-1: CSI compression with two-sided AI/ML model

II-2.3. Training Collaboration at UE-side and NW-side

CSI compression involves a two-sided model where inference processing using AI/ML models is executed on both the UE-side and the NW-side. One of the challenges in considering the two-sided model is how to coordinate training between the UE-side and the NW-side. With this context, several types of training that involve different degrees of collaboration between UE-side and the NW-side have been studied in terms of inter-vendor collaboration complexity, performance, maintainability, and standardization impact. In Release 19, in order to alleviate/resolve the issue related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, the following three directions has been studied.

- Direction A: Sharing parameters/dataset that enables UE-side offline engineering
- Direction B: Sharing NW-side encoder parameter to UE-side for UE-side inference directly with on-device operation
- Direction C: Fully standardized reference model(s) and parameters with specified encoder and/or decoder part

In addition, for Direction A, two types of information sharing have been studied.

- Direction A-1: Encoder parameter exchange, with target CSI
- Direction A-2: Dataset exchange (i.e., target CSI and CSI feedback)

Direction A-1 is the training collaboration option with standardized reference model structure and parameter exchanges with target CSI between NW-side and UE-side. Parameters and target CSI received at the UE-side goes through offline engineering at the UE-side (e.g., UE-side over-the-top (OTT) server). Offline coordination between either NW-side and UE-side or intra-vendor entities is alleviated or not necessary if the model structure can be specified, and parameter exchange is via standardized signaling. One of potential issues is how the mismatch between NW-side data distribution and UE-side data distribution impacts on the performance. For example, AI/ML models trained at NW-side may not reflect data distribution with respect to UE-specific conditions, sush as UE antenna configuration, implementation-specific demodulation algorithms, etc. The AI/ML model working environment of these conditions should be ideally same between the training and inference. We evaluate the performance impact due to the mismatch between NW-side and UE-side data distribution on Direction A-1 in Section 4.1 and showed that Direction A-1 can address the performance impact due to NW / UE data distribution mismatch with respect to UE-specific conditions.

Direction A-2 is the training collaboration option with dataset exchange between NW-side and UE-side. The exchanged dataset includes target CSI and CSI feedback. This option allows each UE/chip set vender of UE-side designs their algorithm with the help of NW-specific information. Since dataset is delivered to UE-side instead of model structure and/or parameters, there is uncertainty on the reference model expression. Therefore, combination with Direction C may be necessary to alleviate the burden of inter-vendor collaboration and/or offline engineering to align model structure between NW-side and UE-side.

Direction B is the training collaboration option with standardized reference model structure and parameter exchanges between NW-side and UE-side. In addition, parameters received at the UE are directly used for inference at the UE without offline engineering, with on-device operations. For Direction B, UE-side model switching only includes the updating parameters, while model training is not needed. On the other hand, this direction may not allow device specific optimization compared to Direction A. The potential issue would be how to provide same working environment, i.e., the parameters/conditions that shall be considered for inference encoder training should be aligned between NW and UE, resulting in potential inter-vendor collaboration effort or more standardization effort.

In Direction C, reference AI/ML model is standardized with the actual implementation of AI/ML models on the UE-side and NW-side based on these reference AI/ML models. It can eliminate the inter-vendor collaboration complexity if feasible for specification. One of issues is reference AI/ML models trained using statistical channel models may not be suitable for real field environments. We evaluate the performance impact due to mismatch between the distribution of the dataset used for reference model training, UE-side data distribution, and NW-side data distribution on Direction C in Section 4.2 and showed that fully specified model in Direction C may have limited performance in the field, but the performance may be improved at least if either side train their part of model using field data.

Table II-2.3-1 are summarized our views on the comparison of directions.

Table II-2.3-1: Comparison of directions

	Inter-vendor collaboration	Performance	Maintainability	Standardization
Direction A-1	Feasible, or complexity is alleviated.	Good	Allowing UE-side and NW-side to develop/update models separately	Feasible
Direction A-2	Complexity is alleviated.	May be worse without backbone structure alignment	Allowing UE-side and NW-side to develop/update models separately	Feasible
Direction B	Large complexity to align the same working environment	Unclear whether the performance impact due to NW-side data distribution and UE-side inference data distribution mismatch can be addressed.	Only NW-side can develop/update the model. Not feasible for UE-side	Feasible but more standardization effort
Direction C	Feasible	Limited compare with other directions	Allowing UE-side and NW-side to develop/update models separately	Feasible

II-2.4. Performance Evaluation

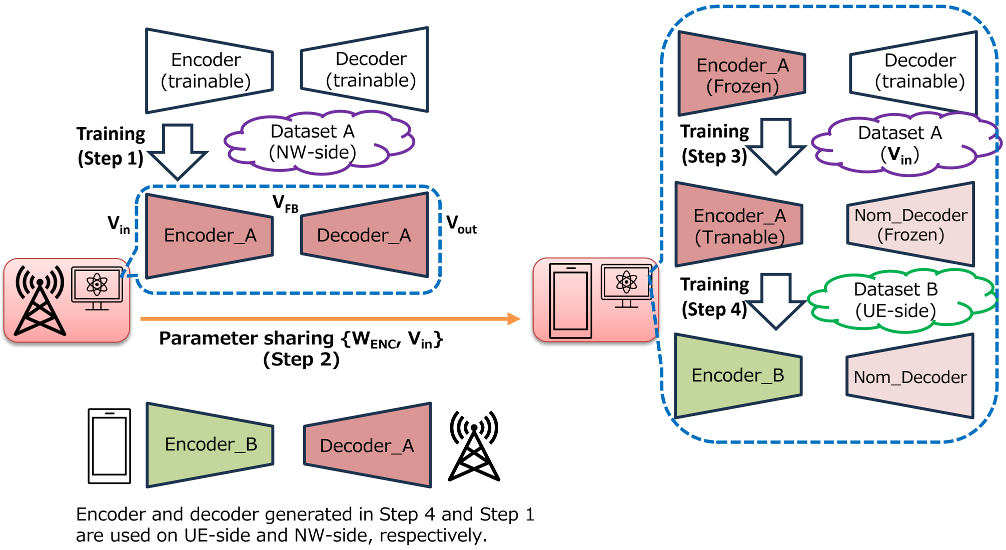
II-2.4.1. Direction A-1

Assuming that the reference model structure is standardized, we consider the following procedure for NW-side and UE-side training as shown in Fig. II-2.4.1-1.

- Step 1: NW-side trains the encoder (which is not used for inference) and decoder jointly.
- Step 2: After NW-side training is finished, NW-side shares UE-side with encoder parameters of the trained encoder model and target CSI (Dataset A) used in the NW-side training.
- Step 3: UE-side first develop a nominal decoder against the exchanged encoder using encoder parameters and target CSI exchanged from NW-side.
- Step 4: UE-side develops actual encoder against the nominal decoder using the target CSI measured at UE-side (Dataset B).

In order to investigate the performance impact on UE-side / NW-side data distribution mismatch with respect to UE-side additional condition, we consider NW-side data (Dataset A) and UE-side data (Dataset B) are mismatched in terms of UE-side antenna configuration. Dataset A and Dataset B are constructed as follows. Detailed parameters for evaluation conditions are shown in Table II-2.4.1-1.

- Dataset A: 3 types of UE antenna configurations, $(M, N, P, M_g, M_g; M_p, N_P)$ = (1, 2, 2, 1, 1, 1, 2), (2, 1, 2, 1, 1, 2, 1), and (2, 2, 1, 1, 1, 2, 2) are assumed.
- Dataset B: Only 1 type of UE antenna configuration, $(M, N, P, M_g, M_g; M_p, N_P)$ = (1, 2, 2, 1, 1, 1, 2) is assumed.



Encoder and decoder generated in Step 4 and Step 1 are used on UE-side and NW-side, respectively.

Fig. II-2.4.1-1: Training collaboration procedure in Direction A-1

Table II-2.4.1-1: Evaluation assumptions

Parameter	**Value**	
Scenario	Dense urban macro	
Frequency range	2 GHz	
Inter-BS distance	200 m	
Channel model	According to TR 38.901	
Antenna setup and port layouts at gNB	32 ports: (8, 8, 2, 1, 1, 2, 8), $(d_H, d_V) = (0.5, 0.8)\lambda$	
Antenna setup and port layouts at UE	4 Rx: For Dataset S and A: (1, 2, 2, 1, 1, 1, 2), (2, 1, 2, 1, 1, 2, 1), and (2, 2, 1, 1, 1, 2, 2), $(d_H, d_V) = (0.5, 0.5)\lambda$ For Dataset B: (1, 2, 2, 1, 1, 1, 2), $(d_H, d_V) = (0.5, 0.5)\lambda$ Note: Antenna configuration is indicated as $(M, N, P, M_g, M_g; M_p, N_P)$, where M and N are the number of vertical, horizontal antenna elements within a panel, P is number of polarizations, M_g is the number of panels in a column, N_g is the number of panels in row; and M_P and N_p are the number of vertical, horizontal TXRUs within a panel and polarization.	
BS antenna height	25 m	
UE antenna height and gain	Follow TR 36.873 [8]	
Numerology	Slot / non-slot	14 OFDM symbol slot
	SCS	15 kHz
Simulation bandwidth	10 MHz	
UE distribution	Dataset S: 80 % indoor (3 km/h), 20 % outdoor (30 km/h) Dataset A, B: 100 % outdoor (30 km/h), various LOS/NLOS ratios (100:0, 40:60, and 20:80) for outdoor UEs are considered.	
Feedback assumption	Ideal	
Channel estimation	Ideal	
Rank number	1	
CSI compression model	Transformer [5, 6]	
Dataset size for training and inference	For Direction A-1 ● 300,900 for NW-side training, UE-side nominal decoder training ● 150,450 for UE-side encoder training ● 39,900 for inference For Direction C ● 300,900 for training Dataset S ● 150,450 for training Dataset A and B ● 39,900 for inference	

We consider the following three cases.

- Case 1A: An encoder-decoder pair is trained in Dataset B. This serves as an upper bound.
- Case 1B: NW-side trains a decoder based on Dataset B, and UE-side trains an encoder based on Dataset B.
- Case 2: NW-side trains a decoder based on Dataset A, and UE-side trains an encoder based on Dataset B.
- Table II-2.4.1-2 shows the squared generalized cosine similarity (SGCS), which represents the similarity between the reconstructed and original CSI. From the comparison between Case 1A and Case 1B, Direction A-1 can achieve almost the same performance as joint training if data distribution between NW-side offline training and UE-side offline training is aligned. From the comparison between Case 1 and Case 2, Case 2 cause performance loss due to data distribution mismatch between Dataset A and Dataset B. On the other hand, in terms of UE-specific condition of antenna layout/configuration, the performance loss is small if Dataset A for NW-side training includes Dataset B.

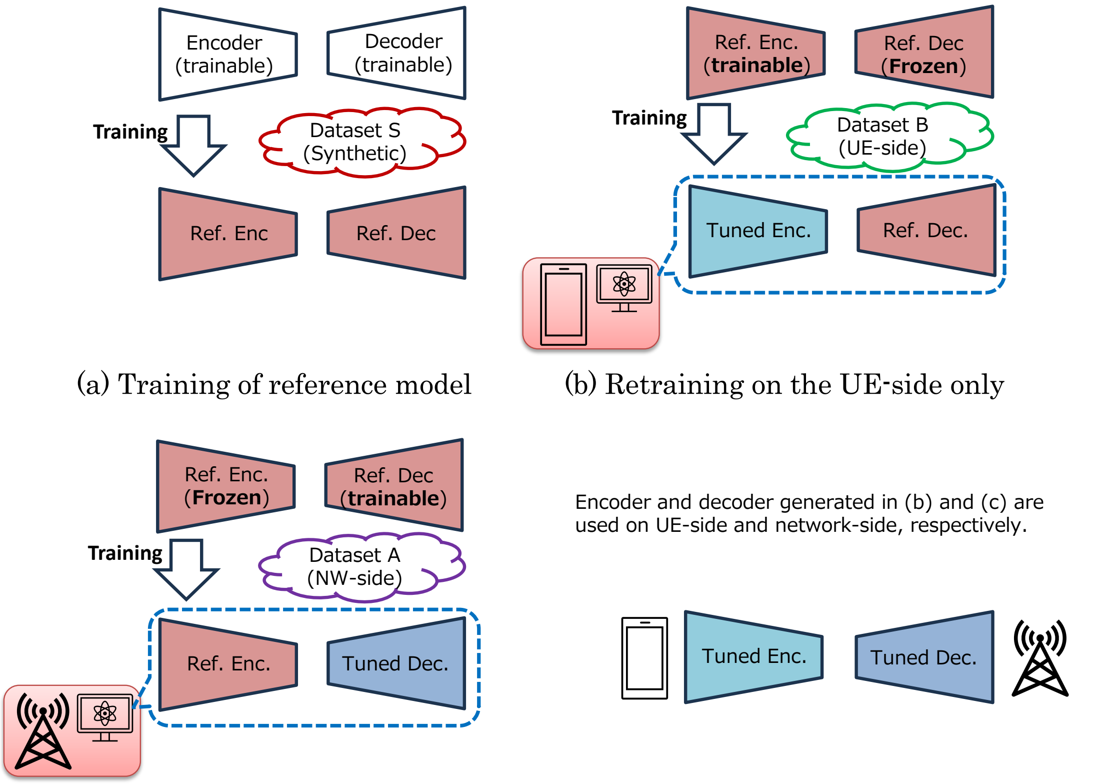
Table II-2.4.1-2: Performance of Direction A-1

Case	Notes	SGCS by inference on Dataset B (Performance loss from upper bound)		
		LOS: 100%	LOS: 40%	LOS: 20%
1A	The encoder-decoder pair is jointly trained based on training Dataset B (upper bound).	0.9428	0.7730	0.7233
1B	NW-side trains a decoder on Dataset B. UE-side trains a nominal decoder and an encoder based on Dataset B.	0.9383 (0.48%)	0.7721 (0.12%)	0.7228 (0.08%)
2	NW-side trains a decoder on Dataset A. UE-side trains a nominal decoder based on Dataset A, and then, UE-side trains encoder based on Dataset B.	0.9407 (0.22%)	0.7680 (0.54%)	0.7155 (1.09%)

II-2.4.2. Direction C

Assuming that the trained reference model is standardized, we compared the impact on CSI reconstruction accuracy in case where the dataset distributions are different between the reference model training and the inference phases. We consider the following four scenarios as shown in Fig. II-2.4.2-1: a) without retraining, i.e., reference model, b) retraining the AI/ML model only on the UE-side, c) retraining AI/ML model only on the NW-side, and d) retraining AI/ML model independently on both the UE-side and NW-sides.

The evaluation conditions are shown in Table II-2.4.1-1. The dataset for reference model training (Dataset S) is based on a dense urban macro scenario with a UE distribution of {80% indoor, 20% outdoor}. In addition, 3 types of UE antenna configurations are assumed. For retraining and inference, Dataset A is used for NW-side retraining and Dataset B is used for UE-side retraining, whose dataset construction is same as in Section 4.1.



(a) Training of reference model (b) Retraining on the UE-side only

Encoder and decoder generated in (b) and (c) are used on UE-side and network-side, respectively.

(c) Retraining on the NW-side only (d) Retraining on the UE-side and NW-sides.

Fig. II-2.4.2-1: Retraining of the reference model

Table II-2.4.2-1: Performance of finetuning encoder and decoder under Direction C

	SGCS by inference on Dataset B (Performance loss from upper bound)		
	LOS: 100%	LOS: 40%	LOS: 20%
The model is trained based on training Dataset B (upper bound).	0.9437	0.7701	0.7234
The specified model is trained based on training Dataset S.	0.9238 (2.12%)	0.7484 (2.83%)	0.6896 (4.67%)
Encoder model is trained against the specified decoder model using Dataset B.	0.9290 (1.56%)	0.7536 (2.16%)	0.6959 (3.81%)
Decoder model is trained against the specified encoder model using Dataset A.	0.9335 (1.08%)	0.7563 (1.80%)	0.6992 (3.34%)
Encoder / decoder model is separately trained against specified decoder / encoder model using Dataset B (at UE-side) and Dataset A (at NW-side). (No inter-vendor collaboration)	0.9320 (1.24%)	0.7541 (2.09%)	0.6968 (3.69%)

Table II-2.4.2-1 shows the SGCS value under Direction C. For comparison, the inference results using AI/ML model trained on the inference dataset is also shown. AI/ML model without retraining shows performance degradation due to the mismatch in dataset distribution. Performance improvements can be observed when the UE-side or

NW-side retrains using datasets that match the inference environment. The performance improvement by retraining on the NW-side only is larger than that by retraining on the UE-side only due to the decoder having more layers and parameters. The improvement is limited when the UE-side and NW-side retrain independently without coordination. This suggests that when retraining models on both the UE-side and NW-side, it is necessary to share the retraining results from the NW-side with the UE-side, resulting in the necessity of Direction A with the combination with Direction C.

II-2.5. Conclusion

We introduced three approaches for training collaboration of AI/ML-based CSI compression using a two-sided model, which will be used as a starting point for studying two-sided AI/ML solutions in 6G. Computer simulations show that performance improvements can be achieved by retraining models on both the UE-side and NW-side while suggesting that sharing retraining results between the NW-side and UE-side is necessary for improving the performance, highlighting the potential need for a combination of Direction A and Direction C. Performance investigation of Direction A-2 is left as our future study.

REFERENCE

[1] RP-221348, “Revised SID: Study on AI/ML for NR air interface,” June 2022.

[2] RP-243244, “Revised WID: Artificial Intelligence (AI) / Machine Learning (ML) for NR Air Interface,” Qualcomm, December 2024.

[3] RP-243245, “New SID: Study on Artificial Intelligence (AI) / Machine Learning (ML) for NR air interface Phase 2,” Qualcomm, December 2024.

[4] RP-234055, New SID: Study on Artificial Intelligence (AI) / Machine Learning (ML) for mobility in NR,” CMCC, December 2023.

[5] C.K. Wen, et. al., “Deep learning for massive MIMO CSI feedback,” IEEE Wireless Commun. Letters, Vol.7, No.5, p.748 – 751, March 2018.

[6] H. Xiao, et. al., “AI enlightens wireless communication: A transformer backbone for CSI feedback,” China communications, June 2022.

[7] 3GPP TR 38.843, “Study on Artificial Intelligence (AI) / Machine Learning (ML) for NR air interface, V2.0.1, January 2024.

[8] 3GPP TR 36.873, “Study on 3D channel model for LTE,” V12.7.0, January 2018.

II-3. Proof-of-concept for AI-native Air Interface Toward 6G

Hiroto Yamamoto, Shuki Wai and Daisei Uchida
NTT Access Network Service Systems Laboratories, NTT Corporation
Atsuya Nakamura, Satoshi Suyama and Huiling Jiang
6G-Tech Department, NTT DOCOMO, INC.
Dani Korpi and Jaime J.L. Quispe
Nokia Bell Labs
Kyungpil Lee and Minsoo Na
SK Telecom Co., Ltd.

Abstract— **For 6G, the use of AI/ML is one of the key technologies and its application to the air interface is being widely considered. This article introduces the proof-of-concept (PoC) for AI-native air interface (AI-AI) which utilizes AI/ML for some functions of the air interface for 6G. The AI-AI PoC is tested in an indoor environment, and the throughput improvement by AI-AI is confirmed. In addition, tests using a channel emulator confirmed that AI-AI can further improve throughput in a high-speed mobile environment.**

II-3.1. Introduction

The application of AI/ML (Artificial Intelligence / Machine Learning) technology to wireless communications have been widely studied, and a vision called AI-native air interface (hereinafter referred to as AI-AI) has been proposed in which AI/ML will be used to optimize the air interface end-to-end [1, 2, 3]. In the 3GPP Release 19 currently under discussion, beam management, positioning, CSI feedback, etc. are being studied as a first step in applying AI/ML to the air interface [4]. AI/ML technology will continue to be one of the key topics in 6G, and AI/ML will be used in a lot of air interface functions.

Nokia, SKT, DOCOMO and NTT are collaborating on the development of AI-AI proof-of-concept (PoC) [5]. We tested the AI-AI PoC, which utilizes AI for some functions of the air interface, in a real environment. This article shows the results of the throughput performance of AI-AI.

II-3.2. AI-AI PoC System

In the proposed AI-AI PoC system (hereinafter referred to as proposed scheme), the transmit constellations and the receiver that handles channel estimation, equalization, and demodulation are jointly learned as shown in Fig. II-3.2-1 [6]. In the training, the simulation data of several propagation environments shown in Table II-3.2-1 are used. In addition, the signal-to-noise ratio (SNR) is randomized between 0 and 20 dB and these random parameters are generated for each frame. The constellations learned by the

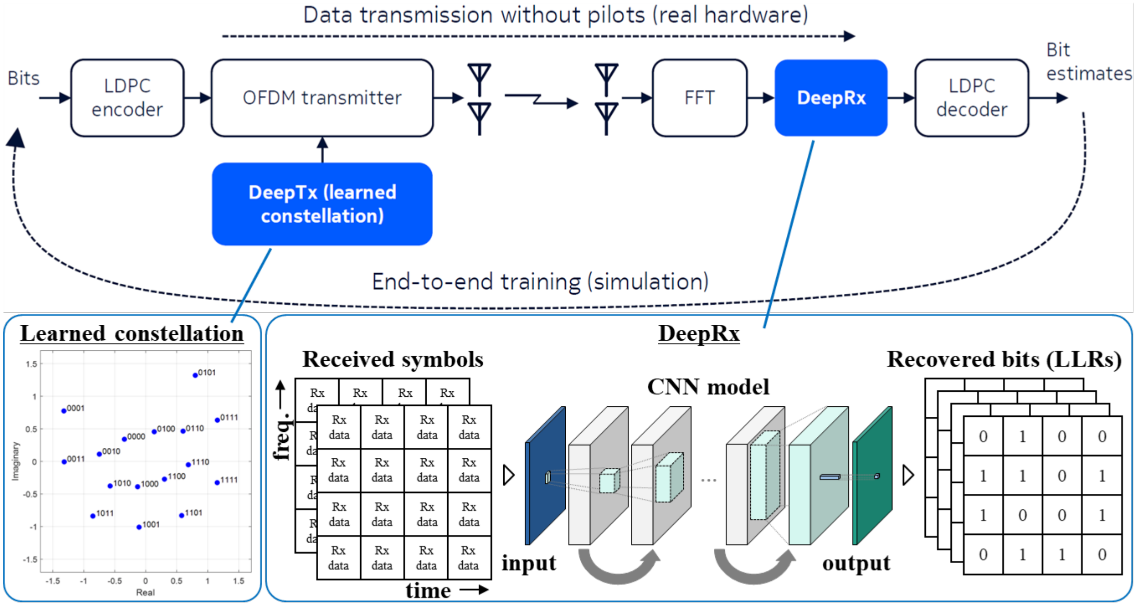


Fig. II-3.2-1. System model of the proposed AI-AI PoC system.

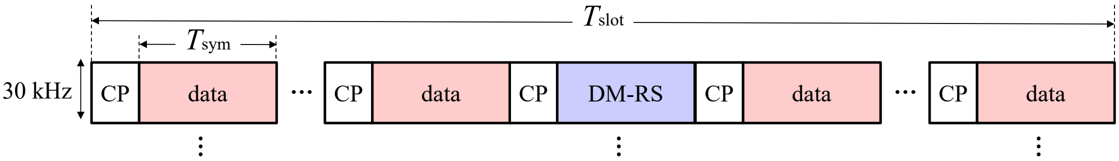


Fig. II-3.2-2. Schematic diagram of the OFDM slots for throughput calculation.

proposed scheme are typically non-uniform patterns as shown in the example in Fig. II-3.2-1. In the receiver, CNN (Convolutional Neural Network) is used to estimate the LLR (Log-Likelihood Ratio) based on the received symbols [7]. In the system, the conventional 5G NR-based scheme (hereinafter referred to as conventional scheme) transmits DM-RSs (DeModulation-Reference Signals) in 2 or 3 OFDM symbols in one slot, while the proposed scheme does not because there is no explicit channel estimation process. Also, neither scheme uses the first symbol for data transmission.

The proposed scheme can transmit data using all available resources, without having to transmit DM-RS, which is expected to improve throughput. In addition, the proposed scheme learns at velocities up to 200 km/h. Therefore, the proposed scheme does not suffer from the degradation of the accuracy of the channel estimate as occurs in conventional scheme, and further performance improvement can be expected in high-speed mobile environments. In the tests of AI-AI PoC, the throughput is measured as calculated by the following equation,

Table II-3.2-1. Simulation parameters for training

Channel model	3GPP TDL-A, TDL-B, TDL-C
Velocity	0~200 km/h
Delay spread	10~500 ns
SNR	0-20 dB

Table II-3.2-2. Specifications of the test

Center frequency	4.8 GHz
Subcarrier spacing	30 kHz
Number of subcarriers	300
MCS index	5~10 (16QAM)

$$C = C_{max} \left(T_{sym} \cdot N_{data}/T_{slot}\right)\left(1 - P_{BLER}\right), \quad (1)$$

$$C_{max} = SCS \cdot N_{subcarrier} \cdot Q_m \cdot R, \quad (2)$$

where T_{sym}, N_{data}, T_{slot} and P_{BLER} are the OFDM symbol duration, the number of data symbol within a slot, the slot duration and block error rate, respectively, as shown in the schematic diagram in Fig. II-3.2-2. C_{max} is the maximum throughput determined by the MCS (Modulation Coding Scheme), and SCS, $N_{subcarrier}$, Q_m and R are the sub-carrier spacing, the number of sub-carriers, the modulation order and the coding rate, respectively. The specifications of the tests are shown in Table II-3.2-2. In the tests, the MCS index for 16QAM specified by 256QAM index table of 3GPP is employed [8].

II-3.3. Indoor Test of AI-AI PoC

This section shows the indoor tests of AI-AI PoC in real environments. Fig. II-3.3-1 shows a schematic diagram and pictures of the AI-AI PoC system. In the test, the baseband processing is implemented in the GPU server and a software-defined radio is used to transmit and receive signals. For the transmitting and receiving antenna omnidirectional antennas are used. Fig. II-3.3-2 and II-3.3-3 show the schematic diagram and picture of the test environment, respectively. In this test, throughputs are measured at the six fixed points indicated by the red dots in Fig. II-3.3-2 in a static test, while in a dynamic test throughputs are measured while moving at walking speed along the measurement routes 1 to 5. The number of DM-RSs is 2 OFDM symbols within a slot and MCS index is 5 in the tests. In the static test, no block errors occurred at any measurement points for both the conventional scheme and the proposed scheme, and consequently the throughput improvement obtained from the ratio of the number of data symbols and block error rate between the proposed and conventional scheme is 18% by excluding DM-RSs. In the dynamic test, the block error rate of the proposed scheme is slightly higher than that of the conventional scheme, but the throughput improvement of the proposed scheme is 6 ~ 16 % as shown in Fig. II-3.3-4.

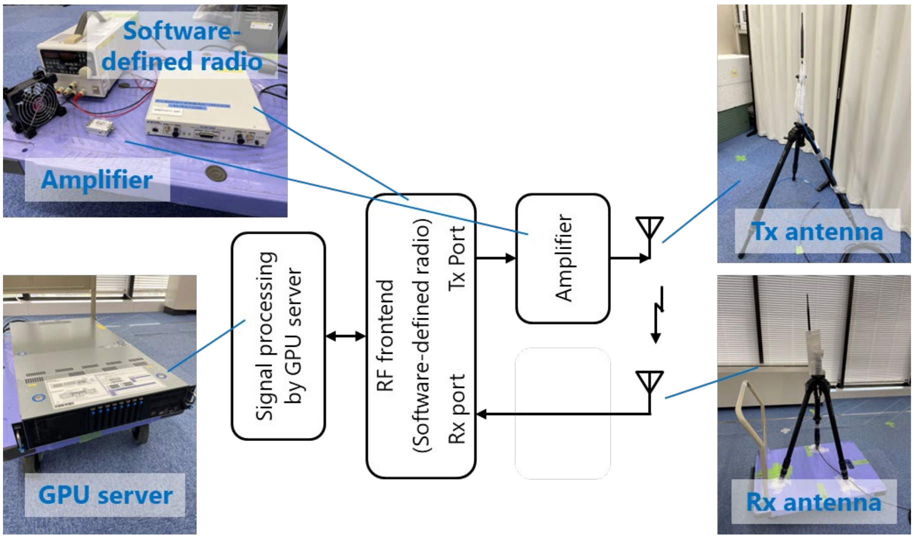


Fig. II-3.3-1. Schematic diagram and pictures of AI-AI PoC system.

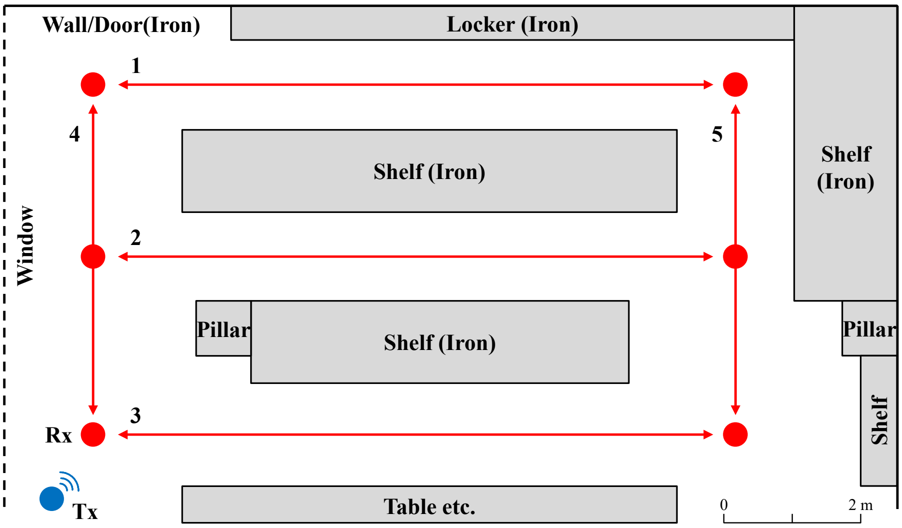


Fig. II-3.3-2. Schematic diagram of the indoor test environment.

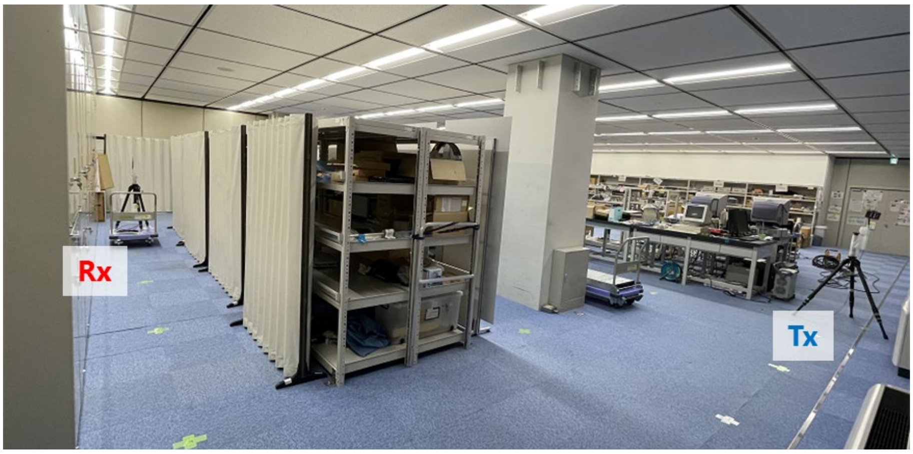


Fig. II-3.3-3. Picture of the indoor test environment.

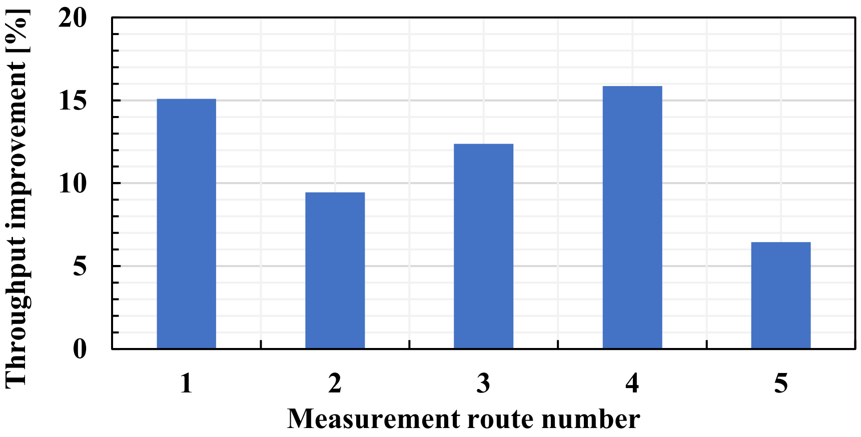


Fig. II-3.3-4. Throughput improvement by the proposed scheme.

II-3.4. Test of AI-AI PoC in Mobile Environments Using a Channel Emulator

This section shows the test of AI-AI PoC in mobile environments using channel emulator which is connected between the transmitter and receiver as shown in the Fig. II-3.4-1. In this test, the performance at speeds of 3 km/h and 120 km/h are measured, and the channel model of WLAN-B is used as it is different from the channel models used for learning [9]. Fig. II-3.4-2 shows the measurement results of throughput versus required SNR if the MCS index for 16QAM is varied. In this figure, the measurement values of throughput and required SNR are plotted when the block error rate (BLER) equals 10^{-1} at each index. For comparison, the characteristics of the conventional scheme in which 3 symbols of DM-RS are inserted are also shown in Fig. II-3.4-2. The figure shows that the proposed scheme has improved throughput compared with the conventional scheme. In particular, when comparing the results of 3 km/h and that of 120 km/h, the performance of conventional scheme deteriorates because it becomes difficult for channel estimation to follow the time variation of the channel, but the performance of the proposed scheme does not show significant degradation. For example, when the SNR is approximately 15 dB at 120 km/h, the proposed scheme operates with MCS index = 8 whereas the conventional scheme operates with MCS index = 7. In this case, the proposed scheme can improve the throughput by about 47% compared with the conventional scheme.

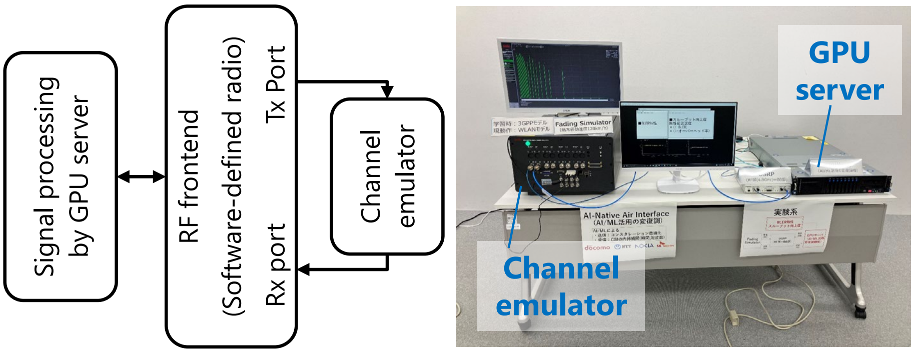

Signal processing by GPU server
RF frontend (Software-defined radio)
Tx Port
Rx port
Channel emulator
GPU server
Channel emulator

Fig. II-3.4-1. Schematic diagram and pictures of AI-AI PoC system with channel emulator.

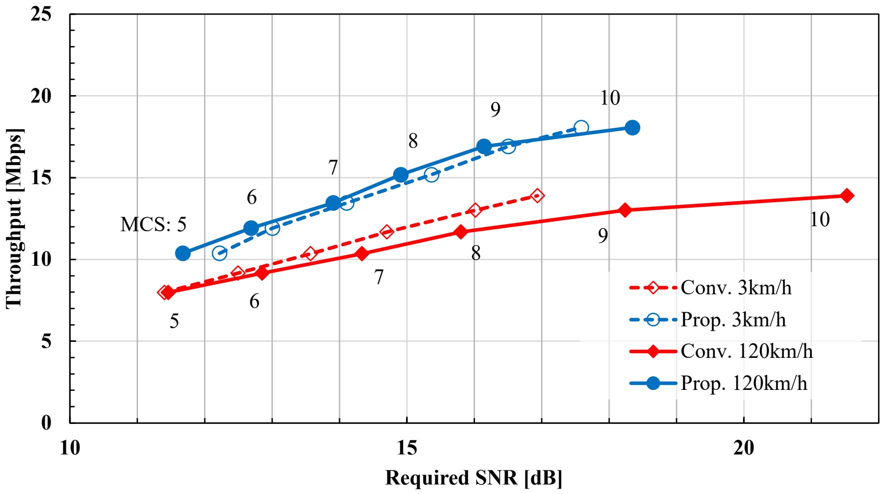


Fig. II-3.4-2. Throughput versus required SNR

II-3.5. Conclusion

In this article, we presented test results for the AI-AI PoC, confirming its effectiveness. In the indoor tests, AI-AI improved the throughput by 6 ~ 18 % compared to the conventional 5G-NR-based scheme. In addition, in the channel emulator tests, it was confirmed that AI-AI can improve the throughput by about 47% in high-speed mobile environments. In the future, we will test the AI-AI PoC in a variety of different environments, including outdoor experiments.

REFERENCE

[1] Jakob Hoydis, Fayçal Ait Aoudia, Alvaro Valcarce, Harish Viswanathan, “Toward a 6G AI-Native Air Interface”, IEEE Communications Magazine, Volume: 59, Issue: 5, May 2021, pp.76-81.

[2] Nokia Bell Labs, "White paper, Toward a 6G AI-Native Air Interface", [Online]. Available: https://www.nokia.com/asset/210299

[3] Mattia Merluzzi, Tamas Borsos, Nandana Rajatheva, et al., "The Hexa-X Project Vision on Artificial Intelligence and Machine Learning-Driven Communication and Computation Co-Design for 6G"

[4] 3GPP Work Item Description, "RP-234039: New WID on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface", 3GPP TSG RAN Meeting #102, Dec. 2023.

[5] NTT DOCOMO, INC., "Press release: 6G Radio-interface Indoor Test Using AI in the 4.8 GHz Band for the First Time in Japan Improves Throughput up to 18%", Nov. 2024. [Online]. Available: https://www.docomo.ne.jp/english/info/media_center/pr/2024/1120_00.html

[6] Dani Korpi, Mikko Honkala, Janne M.J. Huttunen, "Deep Learning-Based Pilotless Spatial Multiplexing", 2023 57th Asilomar Conference on Signals, Systems, and Computers, Oct. 2023.

[7] Mikko Honkala, Dani Korpi, Janne M. J. Huttunen, "DeepRx: Fully Convolutional Deep Learning Receiver", IEEE Transactions on Wireless Communications, Volume: 20, Issue: 6, Jun. 2021, pp.3925-3940.

[8] 3GPP TS38.214, "NR; Physical layer procedures for data", Version: 18.5.0, Jan. 2025.

[9] ETSIEP BRAN 3ER1085B, "Channel Models for HiperLAN/2 in Different Indoor Scenarios", Mar. 1998.

II-4. Neural Network-based Digital Pre-distortion for Wideband Power Amplifiers using DeepShift

Taishi Watanabe, Ohseki Takeo, Issei Kanno
KDDI Research, Inc.

***Abstract*— We present a hardware-efficient neural network-based digital predistortion (DPD) approach for millimeter-wave and terahertz power amplifiers using DeepShift. By replacing multiplications with bitwise-shift and sign operations, our method reduces power consumption by up to 196 times in FPGA and 24 times in 45nm CMOS while maintaining comparable Error Vector Magnitude (EVM) performance. Experimental results with 4.8GHz bandwidth signals demonstrate EVM improvements from 23.96% to 10.23% for millimeter-wave and 55.25% to 20.93% for terahertz power amplifiers. Our DeepShift-based DNN implementation achieves these results with zero multipliers, offering a practical solution for Beyond-5G systems requiring wide-bandwidth nonlinearity compensation.**

II-4.1. Introduction

The evolution of mobile communications toward Beyond-5G and 6G systems demands wider bandwidth operations in millimeter-wave and terahertz bands. At these higher frequencies, power amplifier (PA) nonlinearity becomes a critical challenge, significantly degrading system performance. This degradation is particularly severe in higher frequency bands due to physical device constraints, leading to more pronounced nonlinearity effects. While PA characteristics can be improved through hardware optimization, such improvements often result in reduced power efficiency.

Digital predistortion (DPD) has emerged as a key technique for nonlinearity compensation, applying inverse characteristics to the input signal to counteract PA distortion. Traditionally, DPD has employed polynomial models such as memory polynomials (MP) [1], where PA behavior is expressed as a series of Volterra kernels with different nonlinear orders. These models consider past inputs (memory effects) that influence current output. However, as communication systems expand into wider bandwidths, the complexity of nonlinear distortion increases, making polynomial-based compensation insufficient.

Recently, neural networks (NNs) have been proposed for application in DPD to model complex distortions that occur in wideband systems. Multilayer perceptrons (MLPs) are often used due to their ease of implementation and learning algorithms; based on MLPs, real-valued time-delay neural networks (RVTDNN) [2] have been proposed, which decompose complex signals into real-valued in-phase and orthogonal components and use real-valued learning algorithms RVTDNN takes into account the memory effect of

PA by simultaneously using current and past instantaneous inputs in the input layer. In addition, deep neural networks (DNNs), RVTDNNs with multiple hidden layers, have also been studied to capture more complex nonlinear behavior. While these networks achieve excellent modeling performance, their implementation complexity, particularly the numerous floating-point multiplications, poses significant hardware challenges.

In this paper, the performance of NN-based nonlinear distortion compensation in the millimeter-wave PA and terahertz bands is evaluated experimentally, and efforts to reduce implementation cost by replacing NN multiplication with bit shift and sign operations for the implementation of NN nonlinear distortion compensators are also described.

II-4.2. Neural Network DPD Architecture

This paper employs direct learning [3] as the learning method for neural network-based nonlinear distortion compensation. Direct learning first models actual PA operation and then uses this neural network model to train another neural network model for DPD. We use two architectures: a Real-valued time-delay neural network (RVTDNN) with one hidden layer and a Deep neural network (DNN) with three hidden layers, both using simple fully connected layers. Figure II-4.2-1 shows the RVTDNN architecture. The NN inputs include delayed input signals to the PA, with I/Q signals used as real values. The output predicts and outputs the I/Q signal values of the PA output. Delayed signals are included in the input to model memory effects, where PA output is influenced by past input signals. In our evaluation, hidden layers have twice the number of neurons as the input layer, and tanh is used as the activation function.

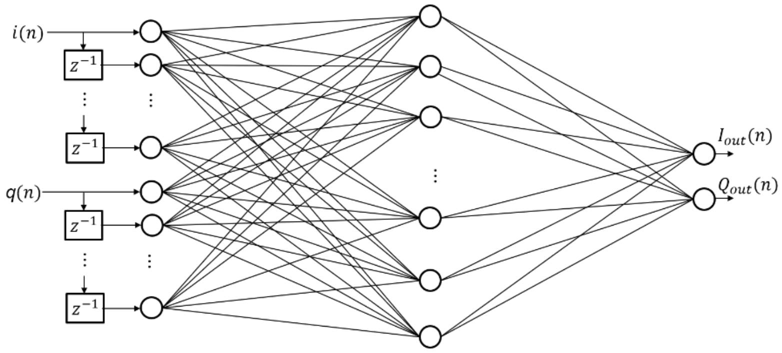


Fig. II-4.2-1. RVTDNN architecture showing the network structure with delayed input signals and I/Q signal processing.

II-4.3. DeepShift

DeepShift is a technique that replaces multiplication operations in neural networks with bit shifts and sign inversions [4]. Figure II-4.3-1 shows an example of operation replacement in DeepShift. The core concept is replacing multiplications with bit shifts by expressing weights as powers of 2. This approach significantly reduces computational complexity and power consumption since bit operations are much simpler to implement in hardware compared to floating-point multiplication. Specifically, bitwise shifts in FIX32 implementation have been shown to reduce power consumption by 24 times and 196 times compared to multiplication in 45nm CMOS technology and FPGA (ZC706), respectively [5]. The authors have demonstrated that DeepShift can be applied to power amplifier modeling while maintaining performance [6].

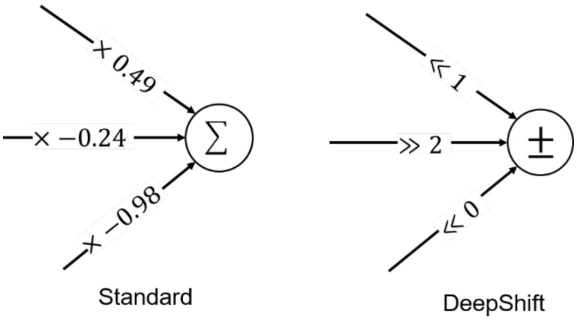


Fig. II-4.3-1. Example of multiplication replacement in DeepShift implementation, demonstrating conversion from standard multiplication to bitwise-shift operations.

II-4.4. Measurement Setup

Experiments were conducted using millimeter-wave and terahertz band PAs. Tables II-4.4-1 and II-4.4-2 show the parameters of the OFDM signals and NN parameters used in the experiments. The memory depth (number of delay taps) used for the NN input signals was 13 for millimeter-wave and 99 for terahertz, with input neuron numbers of 28 and 100, respectively. Performance metrics include the Error Vector Magnitude (EVM) between the actual PA output signal and predicted output.

Table II-4.4-1. OFDM Signal parameters for millimeter-wave and terahertz band OFDM testing, showing key configuration for evaluating DPD performance across different frequency bands.

Center Frequency	37.5 GHz (Millimeter-wave), 261.0 GHz (Terahertz)
Number of Subcarriers	19008
FFT size	32768
Bandwidth	4.8 GHz
Modulation Scheme	QPSK

Table II-4.4-2. Neural network training configuration parameters, detailing optimization settings and DeepShift specifications for both frequency bands.

Stochastic Gradient Descent Method	Adam optimization
Loss Function	Mean Square Error
Mini-batch Size	1024
Number of Epochs	100 (Millimeter-wave), 200 (Terahertz)
Training Symbols	5 OFDM symbols
Learning Rate	Initial value: 0.005 (Millimeter-wave), 0.001 (Terahertz)
DeepShift Weight	Sign part: 1 bit, Bitwise shift part: 4 bit

II-4.5. Experimental Results

Tables II-4.5-1 and II-4.5-2 show the results of applying DeepShift-based DPD to millimeter-wave and terahertz power amplifiers. The results demonstrate that, particularly when using DNN, comparable EVM accuracy to floating-point implementation can be achieved even when applying DeepShift to replace NN multiplications with bit shifts and sign operations. Between DNN and RVTDNN, DNN shows less EVM degradation when applying DeepShift, likely due to its larger scale helping mitigate errors from multiplication replacement.

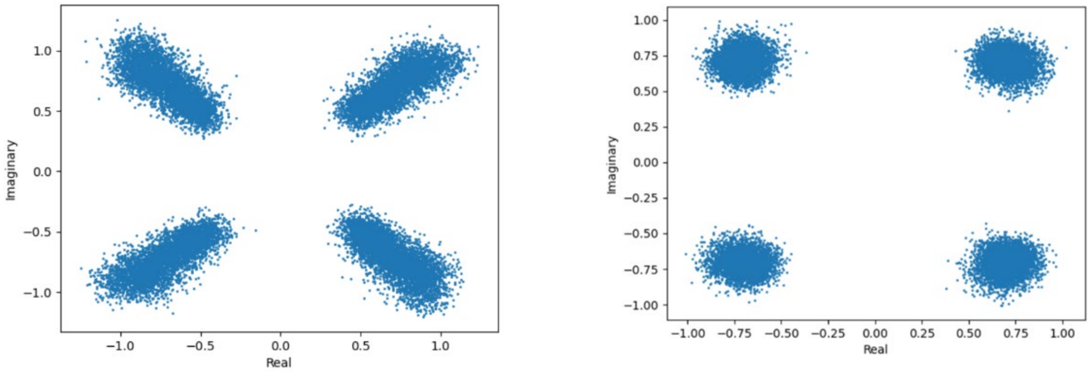
Comparing millimeter-wave and terahertz compensation performance, the performance difference between RVTDNN and DNN is larger for terahertz, indicating that terahertz requires more sophisticated models due to more complex distortion. The AM-AM characteristics show that compensation suppresses characteristic spreading for both millimeter-wave and terahertz, indicating successful mitigation of memory effects. The power spectra show improved flatness across frequency bands after distortion compensation.

Table II-4.5-1. Performance comparison of nonlinear distortion compensation for millimeter-wave power amplifier, showing operation counts and EVM improvement for different model architectures.

Model type	EVM [%]	Multiplication	Bitwise shift& sign	Add	Activation
Without DPD	23.96				
RVTDNN	12.37	1680	0	1680	56
RVTDNN (DeepShift)	14.65	0	1680	1680	56
DNN	10.13	7952	0	7952	168
DNN (DeepShift)	10.23	0	7952	7952	168

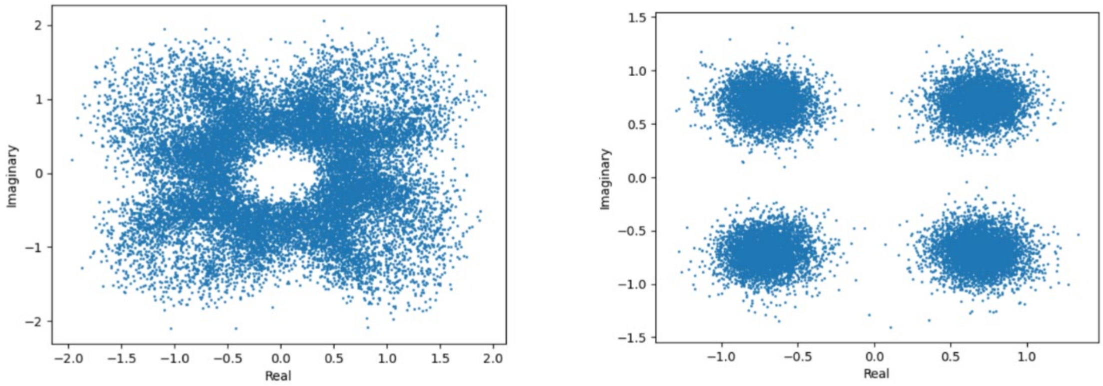
Table II-4.5-2. Performance comparison of nonlinear distortion compensation for terahertz power amplifier, demonstrating computational efficiency and EVM improvement across different architectures.

Model type	EVM [%]	Multiplication	Bitwise shift& sign	Add	Activation
Without DPD	55.25				
RVTDNN	29.92	20400	0	20400	200
RVTDNN (DeepShift)	29.96	0	20400	20400	200
DNN	20.74	100400	0	100400	600
DNN (DeepShift)	20.93	0	100400	100400	600



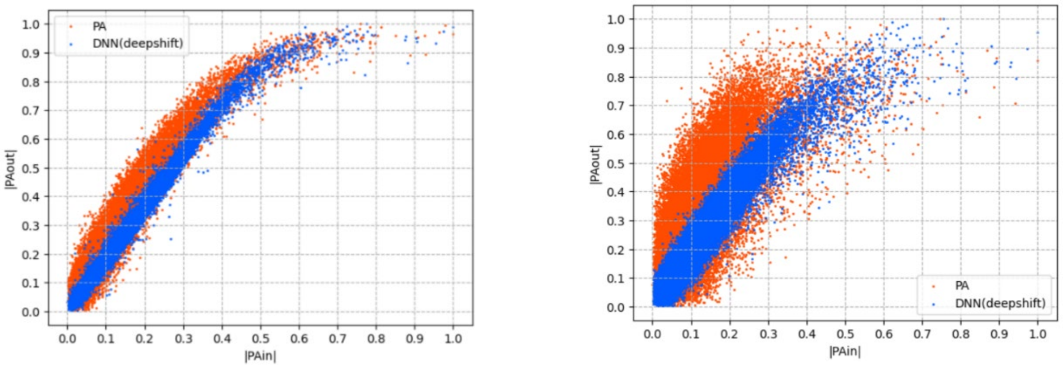
(a) Without DPD (b) DNN (DeepShift)

Fig. II-4.5-1. Constellation diagrams for millimeter-wave PA (a) without DPD and (b) with DNN using DeepShift, showing improvement in signal quality.



(a) Without DPD (b) DNN (DeepShift)

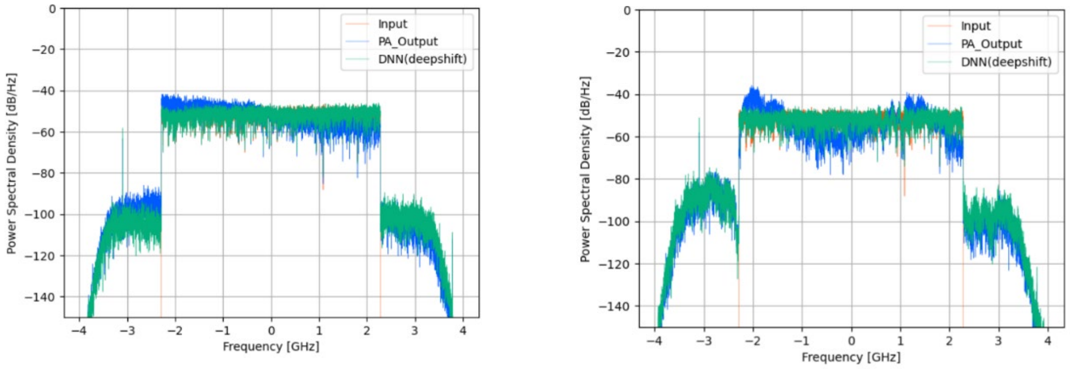
Fig. II-4.5-2. Constellation diagrams for THz PA (a) without DPD and (b) with DNN using DeepShift, showing improvement in signal quality.



(a) Millimeter-wave

(b) Terahertz

Fig. II-4.5-3. AM-AM characteristics comparison of input signal, PA output, and corrected output with DeepShift applied.



(a) Millimeter-wave

(b) Terahertz

Fig. II-4.5-4. Power spectral density comparison of input signal, PA output, and corrected output with DeepShift applied.

II-4.6. Conclusion

This paper demonstrated nonlinear distortion compensation for millimeter-wave and terahertz band power amplifiers with 4.8GHz bandwidth signals. To reduce hardware implementation costs, we applied DeepShift, replacing neural network multiplications with low-cost bitwise shifts and sign operations. Experimental results confirmed that DeepShift implementation achieved EVM accuracy comparable to floating-point compensation. Future work includes evaluating DeepShift application to more complex neural network models such as RNNs.

Acknowledgements

These research results were obtained from the commissioned research (No.00401) by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCE

[1] W. Huadong, B. Jingfu, W. Zhengde, and H. Jingfu, "An Memory polynomial model for power amplifiers," in 2008 International Conference on Communications, Circuits and Systems, May 2008, pp. 1346–1349.

[2] T. Liu, S. Boumaiza, and F. M. Ghannouchi, "Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks," IEEE Transactions on Microwave Theory and Techniques, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.

[3] G. Paryanti, H. Faig, L. Rokach, and D. Sadot, "A Direct Learning Approach for Neural Network Based Pre-Distortion for Coherent Nonlinear Optical Transmitter," Journal of Lightwave Technology, vol. 38, no. 15, pp. 3883–3896, Aug. 2020.

[4] M. Elhoushi, Z. Chen, F. Shafiq, Y. H. Tian, and J. Y. Li, "DeepShift: Towards Multiplication-Less Neural Networks," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2021, pp. 2359–2368. doi: 10.1109/CVPRW53098.2021.00268.

[5] H. You, B. Li, S. Huihong, Y. Fu, and Y. Lin, "ShiftAddNAS: Hardware-inspired search for more accurate and efficient neural networks," in International Conference on Machine Learning. PMLR, 2022, pp. 25 566–25 580.

[6] T. Watanabe, T. Ohseki, and Y. Amano, "Digital Predistortion of RF Power Amplifiers using DeepShift," in 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom: IEEE, Mar. 2023, pp. 1–6.

II-5. AI Calibration Network under Hardware Limitations

Keita Kuriyama, Kentaro Tanaka, Hitoshi Hasegawa, Toshifumi Miyagi
NTT Access Network Service Systems Laboratories, NTT Corporation
Satoshi Suyama, Huiling Jiang, Atsuya Nakamura
6G-Tech Department, NTT DOCOMO, INC.

***Abstract*— In the sixth-generation mobile communication system (6G) era, high-frequency bands (e.g., sub-terahertz (sub-THz) bands) show promise for achieving extremely high-speed and high-capacity communications. However, it is difficult to ensure radio frequency (RF) circuit quality in high-frequency bands compared to lower frequency bands below mmWave, and utilizing higher performance and higher quality circuits lead to higher costs. Considering the increasing popularity of the high-frequency bands, it is vital to achieve low cost while simultaneously ensuring communication quality. It is also necessary to optimize the communication quality, cost, and power consumption of the whole radio access network (RAN) by considering the diversified topology. This chapter presents a conceptual overview of our proposed artificial intelligence (AI) device calibration and the AI calibration network to optimize communication quality and calibration cost based on AI. Further, as a basic examination of AI device calibration, we demonstrate a compensation technology for RF impairments based on a deep neural network (DNN).**

II-5.1. Introduction

One of the requirements for sixth-generation mobile communication systems (6G) is an extremely high data rate and capacity. Radio access technologies (RAT) to provide data rates over 100 Gbps are being discussed as a potential way of meeting this requirement. To achieve 100 Gbps, exploiting higher frequency bands between 100 GHz and 300 GHz with a wider bandwidth than 5G (e.g., sub-terahertz (sub-THz) bands) is a promising approach. However, to utilize the sub-THz bands in 6G, similar to the case when introducing the millimeter-wave band in 5G, there are many technical issues that need to be resolved by the 2030s. These technical issues are diverse and exist mainly in four areas: radio propagation, radio frequency (RF) devices, modulation and demodulation schemes, and air interfaces, which are summarized in detail for each area in [1].

Focusing on the technical issues in RF devices, the characteristics of the RF devices mainly depend on the frequency band and the signal bandwidth, and it is difficult to ensure the same circuit characteristics in the high-frequency bands as those in the low frequency [2]. Some RF impairments, such as frequency selectivity, IQ imbalance, direct current (DC) offset, carrier leakage, phase noise, and nonlinear distortion, have become increasingly pressing as implementation challenges because they can cause degradation of the communications quality. While performance enhancement and high integration of the RF devices are required, for the popularization of high-frequency bands, the RF devices must be manufactured with a level of accuracy and a cost that enables usage in 6G commercial services. Cost reduction can be achieved by allowing the use of low-quality devices, but a calibration scheme by digital signal processing (DSP) is necessary for ensuring the communication quality. Although a partial compensation technique for the RF impairments by DSP has been proposed, this technique requires designing an optimal digital calibration for high-frequency bands, where the influence of each RF impairment is both large and mixed. In recent years, technology that compensates for multiple RF impairments by utilizing artificial intelligence (AI) such as deep neural networks (DNN) has attracted attention [3].

The performance of the RF device and the resources (e.g., processor capability and power consumption) available for DSP are different for each wireless device. Therefore, it is desirable that the cost of digital calibration is dynamically optimized in accordance with the constraints of the available hardware and the required quality from the applications. In addition, it is necessary to optimize the communication quality, cost, and power consumption of the whole radio access network (RAN) by considering the diversified topology. In this chapter, we present our concept of AI device calibration and the AI calibration network [4], which utilizes AI to optimize the communication quality and the calibration cost. Also, as a basic technology of AI device calibration, a demodulation technology [5] utilizing DNN is introduced.

II-5.2. Concept of AI Calibration

II-5.2.1. AI Device Calibration

Fig. II-5.2.1-1 shows the concept of AI device calibration. The required specifications for the equipment differ between base station (BS) and user equipment (UE), which means the characteristics of the RF circuit, the capability of the processor, and the allowable power consumption for DSP are also different from each device. On the other hand, regarding the communication quality, the total communication quality that can be observed end-to-end only needs to meet the requirements. In other words, points such as calibrating within the transmitting station to send a high-quality signal, calibrating only at the receiving station, setting the proportion of each calibration processing load, etc. can be freely designed. In AI device calibration, the existence of digital calibration in each station, the compensation scheme to be used, and the level of compensation accuracy are dynamically controlled in accordance with the constraint of available hardware and the demands of the communication quality. As a case study of a compensation scheme, we introduce a DNN demodulator in Section 3.1.

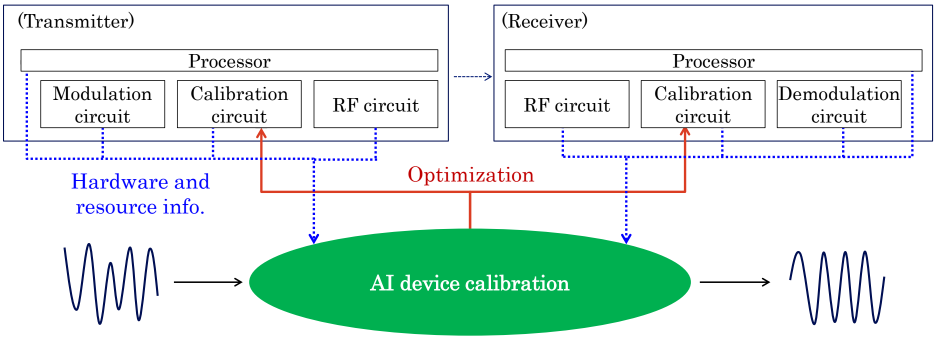


Fig. II-5.2.1-1. Concept of AI device calibration.

II-5.2.2. AI Calibration Network

In the 6G era, many devices will be connected to the RAN through various frequency bands, so the topology of the RAN is expected to diversify. Therefore, new radio network topology (NRNT) has been investigated to improve the performance of RAN for 5G Evolution and 6G [6]. In NRNT, the topology changes dynamically in accordance with the environment, situation, and requirements based on various key performance indicators (KPIs). With sub-THz bands, optimization in the whole of RAN including the relay station (RS) is required, since the impact on KPIs such as equipment cost and power consumption seems to be large.

Fig. II-5.2.2-1 shows the concept of the AI calibration network. The AI device calibration described in Section II-5.2.1 is extended to the entire RAN. The AI calibration network collects the resource information such as the performance of the RF devices of each equipment, processor capability, and power consumption available for DSP. The calibration cost of each device is controlled in accordance with the constraint of available hardware and the requirement of communication quality. In addition, appropriate route selection is performed for the optimization of the whole network.

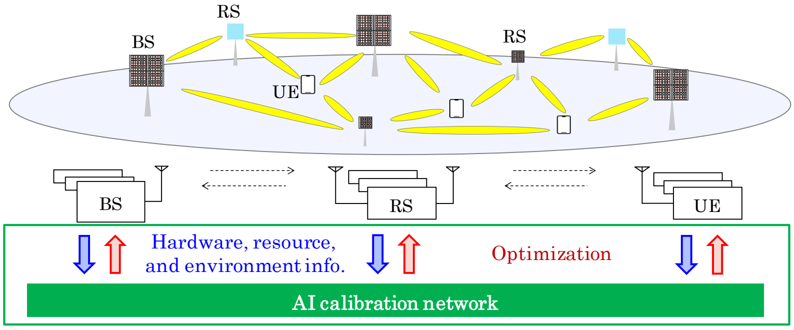


Fig. II-5.2.2-1. Concept of AI calibration network.

II-5.3. Compensation Technique based on AI

II-5.3.1. Deep Neural Network Demodulator

In this section, we introduce a technique to compensate for multiple RF impairments based on DNN. Fig. II-5.3.1-1 shows the system model of our DNN demodulator. Assuming single carrier-(SC) frequency domain equalization (FDE) transmission with nonlinear distortion of the amplifier and IQ imbalance of the IQ modulator/demodulator, bit detection by DNN is performed for the received symbols after FDE. The FDE output is divided into an IQ real valued sequence after IFFT processing and passed to the input layer of the DNN. The DNN outputs a vector consisting of values from 0 to 1 via the sigmoid function. The DNN is trained to minimize the root mean square error (RMSE) of the training data and the output data. Round processing is performed after the DNN, and the bit string corresponding to the index of the input data is output.

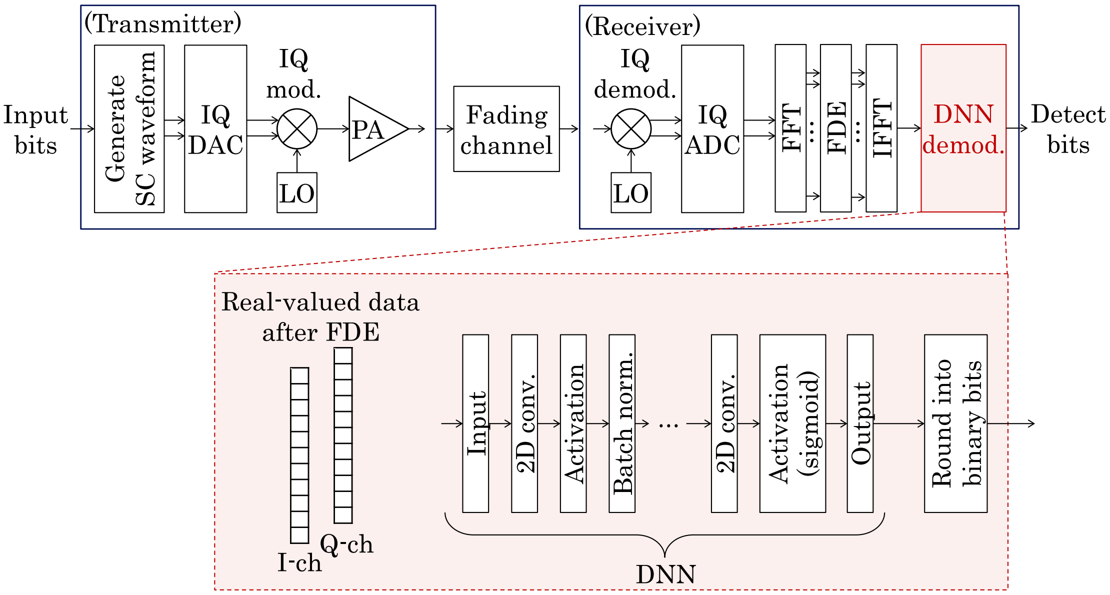


Fig. II-5.3.1-1. System model of DNN demodulator.

II-5.3.2. Numerical Results

In this section, we evaluate the effectiveness of the DNN demodulator. Table II-5.3.2-1 lists the parameters utilized during learning and validation. The nonlinearity of the power amplifier utilizes a Rapp model with the input back off (IBO) of 2 dB. The IQ imbalance of the quadrature modulator and demodulator is 1 dB in amplitude and 5° in phase. The average signal to noise power ratio (SNR) is 30 dB while learning the DNN. The fading channel assumes a static environment and utilizes common values in the learning and test data. We use 16-QAM and 64-QAM for the modulation scheme.

Fig. II-5.3.2-1 shows the bit error rate (BER) performance of the DNN demodulator, w/o compensation, and w/o RF impairments of nonlinear distortion and IQ imbalance. In w/o compensation, hard decision demodulation is performed after IFFT. Compared to w/o compensation, the DNN demodulator improves the BER performance and approaches the characteristics of the case without RF impairments. These results demonstrate that the communication quality can be improved by using the DNN demodulator for RF impairments.

Table II-5.3.2-1. Simulation parameters for training and validation.

Parameter	Value
Modulation scheme	16-QAM, 64-QAM
PA Nonlinearity	Rapp model[7], p=2, IBO=2 dB
Tx and Rx IQ imbalance	Amplitude: 1 dB Phase: 5°
Channel model	Static, exponential
CIR length	4
Average SNR	Training: 30 dB Validation: 10 dB – 30 dB
FFT size	64

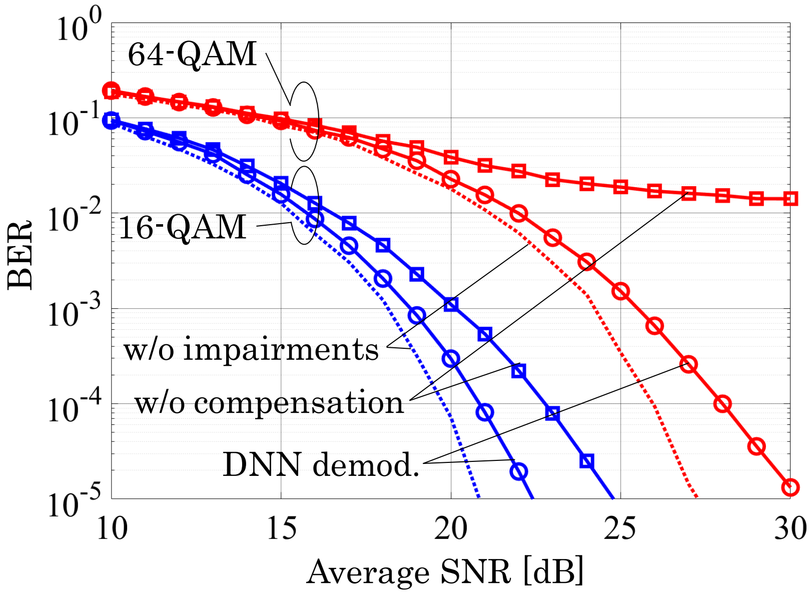


Fig. II-5.3.2-1. BER performance with PA nonlinearity, IQ imbalance, and fading channel.

II-5.4. Conclusion

To popularize the use of high-frequency bands, it is desirable to optimize the trade-off between communication quality and cost reduction in accordance with the application requirements. We introduced a conceptual overview of AI device calibration as an optimization technology that dynamically controls the calibration cost in accordance with the constraint of available hardware and the demand of the communication quality. We also proposed an AI calibration network to optimize the device calibration cost across the RAN, which is an important KPI in the sub-THz bands. Our findings demonstrated the effectiveness of the RF impairment compensation technology based on DNN as a basic technology for AI device calibration.

REFERENCE

[1] H. Fukuzono, S. Suyama, D. Lee, D. Uchida, T. Okuyama, M. Iwabuchi, J. Mashino, and Y. Kishiyama, "Issues of Sub-Terahertz-Band Radio Access Technologies for 6G," IEICE Tech. Rep., vol. 121, no. 302, RCS2021-180, pp. 30–34, Dec. 2021.

[2] U. Gustavsson et al., "Implementation Challenges and Opportunities in Beyond-5G and 6G Communication," IEEE Journal of Microwaves, vol. 1, no. 1, pp. 86–100, Jan. 2021. DOI: 10.1109/JMW.2020.3034648

[3] J. Pihlajasalo, D. Korpi, T. Riihonen, J. Talvitie, M. A. Uusitalo, and M. Valkama, "Detection of Impaired OFDM Waveforms Using Deep Learning Receiver," 2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC), Oulu, Finland, 2022, pp. 1–5. DOI: 10.1109/SPAWC51304.2022.9834021

[4] K. Tanaka, K. Kuriyama, H. Hasegawa, T. Miyagi, S. Suyama, and Takayuki Yamada, "A study of calibration techniques for the use of high frequency bands," in Proc. IEICE Society Conf., B-5-46, Sept. 2023.

[5] K. Kuriyama, K. Tanaka, H. Hasegawa, T. Miyagi, S. Suyama, and Takayuki Yamada, "A Study of Compensation for RF Impairments Based on DNN in SC-FDE Transmission with Higher Frequency Bands," in Proc. IEICE Society Conf., B-5-46, Sept. 2023.

[6] M. Iwabuchi, S. Suyama, T. Arai, M. Nakamura, K. Goto, R. Ohmiya, D. Uchida, T. Yamada, and T. Ogawa, "Concept and Issues of New Radio Network Topology for 5G Evolution & 6G," IEICE Tech. Rep., vol. 122, no. 235, RCS2022-148, pp. 101–106, Oct. 2022.

[7] T. Schenk, RF imperfections in high-rate wireless, Springer

II-6. Performance Requirements and Evaluation Methodology for AI and Communication in 6G

Takashi KOSHIMIZU
Huawei Technologies Japan
Jiang WANG
Huawei Technologies

***Abstract*— The ITU-R has defined "AI and Communication" as one of the six usage scenarios for 6G systems. However, its KPIs and minimum requirements are yet to be defined. This paper describes the "AI and Communication" scenario and the typical AI services in 6G. It also introduces general principles for performance definition, and detailed performance indicators with the requirements. Then, this provides an evaluation methodology for the proposed performance indicators, along with an example of evaluation procedures.**

II-6.1. Introduction

With rapid development of AI technologies, it becomes an essential feature in industries and society. Mobile systems will also revolutionarily be evolved as a unified infrastructure that integrates communication and AI that delivers ubiquitous AI services in the 6G era.

This paper aims to provide guidelines for designing 6G systems and ensure users receive guaranteed AI services. Specifically, it describes the "AI and Communication" scenario defined in the IMT-2030. Next, this summarizes the current status of the performance indicators, then, introduces the design principles, the proposed qualitative and quantitative performance requirement definitions. Finally, it provides the corresponding evaluation methodology with examples.

II-6.2. AI and Communication

6G aims to make intelligence inclusive by providing AIaaS. By utilizing the data and resources of distributed intelligent terminals, 6G will provide AI model training services, made possible through local training at distributed terminals and model interaction between them over the network. 6G can also provide high-accuracy inference services for resource-constrained terminals by joint scheduling

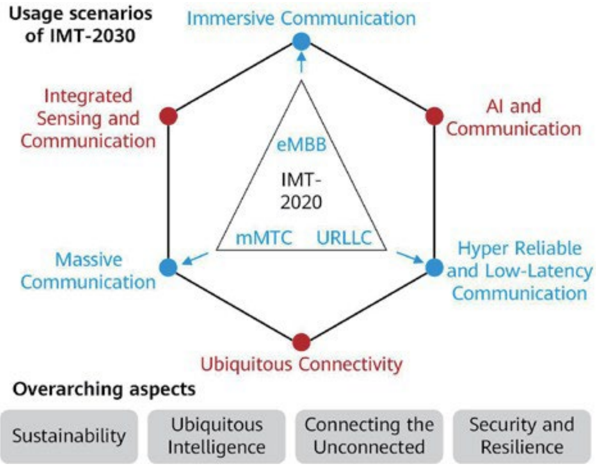


Fig. II-6.2.1-1, Usage scenario of IMT-2030 of communication and AI resources. This will drive AIaaS to become a typical application scenario of 6G.

II-6.2.1. AI and Communication in the IMT-2030 Framework

To facilitate the development of IMT-2030 and beyond, the ITU-R WP-5D approved a new framework [1]. The six usage scenarios identified by the ITU-R are shown in Fig. II-6.2.1-1. To support the new usage scenarios, IMT-2030 includes AI- and sensing-related capabilities, as listed in Table II-6.2.1-1. The "AI and Communication" usage scenario would require high area traffic capacity and user experienced data rates, as well as low latency and high reliability. In addition to the communication aspects, a set of new capabilities related to the integration of AI functionalities is expected, including data acquisition, preparation and processing from different sources, distributed AI model training, model sharing and distributed inference across IMT systems.

Table II-6.2.1-1 Capabilities of IMT-2030

Enhanced Capabilities	IMT-2020	IMT-2030
Peak data rate (Gbps)	20/10 for DL/UL	e.g., 50, 100, 200
User experienced data rate (Mbps)	100/50 for DL/UL	e.g., 300, 500
Spectrum efficiency (bps/Hz)	(Peak) 30/15 for DL/UL	e.g., x1.5, x3
Area traffic capacity (Mbps/m^2)	10	e.g., 30, 50
Connection density (devices/km^2)	10^6	10^6-10^8
Mobility (km/h)	500	500-1000
Latency (ms)	1	0.1-1
Reliability	$1 - 10^{-5}$	$1 - 10^{-5}$ to $1 - 10^{-7}$
New Capabilities of IMT-2030		**Value**
Coverage		*TBD*
Sensing-related capabilities		*TBD*
AI-related capabilities		*TBD*
Sustainability		*TBD*
Positioning (cm)		1-10

II-6.2.2. Typical Services in the "AI and Communication" Scenario

IMT-2030 will efficiently support AI applications in an end-to-end manner, connecting distributed intelligence to provide ubiquitous AI services. It required to build a distributed and efficient AI service platform by utilizing the connection, data, and model resources in the network.

II-6.2.2.1 An Exemplary AI Application Served by IMT-2030, collaborative robots as in Fig. II-6.2.2-1 are widely recognized as a future 6G application scenario that requires AI services with low latency and high learning and inference accuracy. In this use case, multiple robots work together to accomplish complex tasks in an industrial environment. Through local vision or control models, the robots will be able to detect objects from the sensed images and plan the path trajectory with corresponding control decisions for the subtasks. These robots can cooperate with each other over the network to improve the performance of local models via collaborative training.

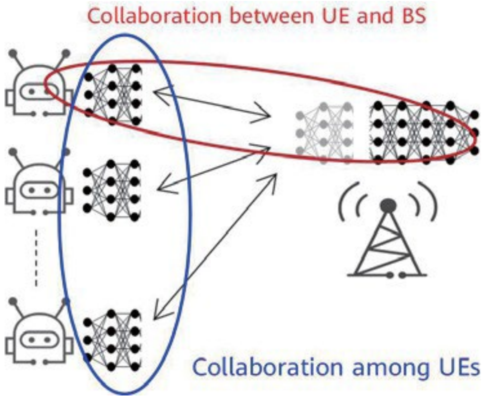


Fig. II-6.2.2-1 AI applications for collaborative robots

II-6-2.2.2 Model Inference Service

AI model inference is a fundamental function for AI applications. It takes inputs, runs the AI models, and produces the expected outputs. Through ubiquitous connectivity, the 6G network with native intelligence could provide real-time model inference capabilities that meet different requirements. Fig. II-6.2.2-2 illustrates a typical AI model inference service. In this service, a large model may be split into two parts, which are deployed on the network and user sides and work together.

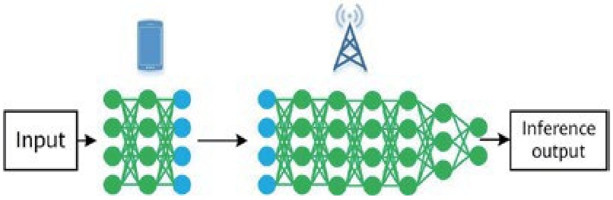


Fig. II-6.2.2-2 AI model inference service

I-6-2.2.3 Model Training Service

AI model training is key for obtaining a model with high accuracy. In the large-scale distributed AI model training service, the network serves as a management platform to provide high-speed data channels and efficient scheduling mechanisms for exchanging data or model parameters between distributed terminals. Fig. II-6.2.2-3 illustrates a typical distributed training service. In each round, the distributed terminals use local data to train models locally and upload the updated models to the network for aggregation.

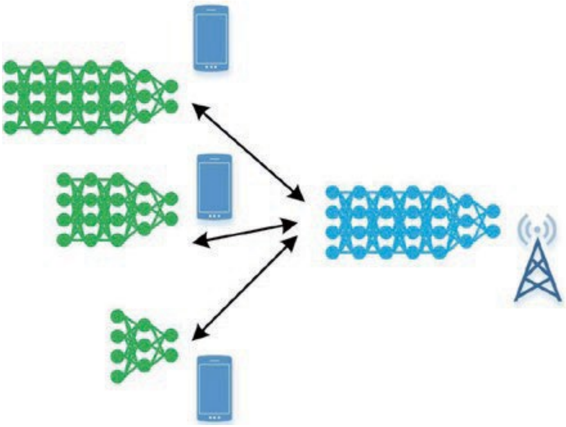


Fig.II-6.2.2-3 Distributed AI model training service

II-6.3. Performance Requirements for the "AI and Communication" Scenario

System design is driven primarily by performance requirements, which evolve or revolutionize each generation of mobile systems. AI services not only involve transmissions, but also include AI-related resources, meaning that AI model learning/inference accuracy and latency are the KPIs. From an AI perspective, the 6G network should support large-scale distributed learning and real-time inference. The 6G network should consider both AI and communication in an integrated manner from the beginning.

II-6.3.1. Current Status

Conventional mobile communication systems have mainly provided communication services. AI&ML features have been studied from Rel-18 such as in TR 22.874 [2]. Various applications have also been defined, however, all the AI/ML operations are expected to be executed in cloud servers. 6G requires new AI-related capabilities that expected to be introduced beyond communication, e.g., supporting AI services. Studies [3, 4], the China IMT-2030 Promotion Group and Hexa-X, both identify AI services provided by 6G as key factors. However, the performance indicators are not illustrated clearly, and no details are defined for the requirements and evaluation methodology toward 6G. The computer science community has defined some AI training and inference KPIs to evaluate the capabilities e.g., MLPerf benchmark [5]. However, these KPIs are used to measure the hardware or software capabilities in a centralized way, it cannot be used to measure the capabilities of distributed AI services.

II-6.3.2. Principles for Performance Definition for AI and Communication in 6G

6G AIaaS will provide various AI capabilities that adapt to different application scenarios. Accordingly, 6G AIaaS needs to consider integrating communications capabilities and AI capabilities in order to build comprehensive performance indicators and evaluation methods.

The main principles of performance definition for AI-related capabilities can be listed;

- End-to-end AI capabilities. AI services should use end-to-end performance as indicators in order to guarantee user-experienced service quality. The AI service quality depends on both communication and AI capabilities.
- Typical services. The IMT-2030 system is the key to realizing ubiquitous intelligence. By utilizing the AI capabilities within the network, this system should provide a platform for large-scale distributed model training and unified high-accuracy model inference.

- Core performance. The goal of AI and communication integration is to enable AI services efficiently, including model training and real-time high-accuracy model inference. To ensure that AIaaS is acceptable to billions of users, it is crucial to focus on the key factors that impact the user experience.

II-6.3.3. Proposed Performance Requirements for AI and Communication in 6G

The KPIs for AI and communication are defined from the perspective of services (including AI model training and inference) provided by 6G networks. The performance of such services depends on the AI model capabilities provided by the system's AI resources and the communication capabilities. Expected qualitative and three quantitative requirements are described below.

- AI service functionality requirements

The functionality requirements for AI-related capabilities are that the candidate radio interface technologies or sets of radio interface technologies shall have mechanisms and/or signaling related to the functionalities, e.g., distributed data processing, distributed learning, AI computing, AI model execution, and AI model inference.

- AI service accuracy (or AI service quality)

AI service accuracy is defined as the accuracy of the AI inference/learning service. Specifically, it is the degree to which the outputs from the AI service are the same as the true values for the given inputs within the given service latency requirements. For a given AI task, the AI service accuracy depends on the task characteristics, AI model deployment method, and AI-related data transmissions.

- AI service latency

AI service latency is defined as the time taken from the start to the end of the AI inference/learning service. It is the sum of the communication time for AI-related data transmissions and the processing time of the AI model, where the processing time depends on the devices and implementations.

- AI service density

AI service density is defined as the number of AI services that meet given AI service accuracy and AI service latency requirements supported by the network simultaneously per unit area. It is a system capacity indicator of the IMT-2030 system. For different application requirements (i.e., accuracy or latency), the system can support different AI service densities.

II-6.4. Evaluation Methodology and Example

Service performance is determined by both communication and AI resources and should therefore be evaluated with certain communication and AI assumptions. This section will describe the evaluation methodology first and then present an example with detailed assumptions and results.

II-6.4.1. Evaluation Methodology

The performance requirements can be derived from two essential KPIs, namely, AI service accuracy and latency. AI service accuracy is defined as the degree to which the outputs from the AI service are the same as the true values for the given inputs. AI service latency is defined as the sum of AI model processing time and data transmission time, which also depends on both the AI model and AI-related data or model transmissions. The performance evaluation can follow the service procedures. Fig. II-6.4.1-1 shows proposed AI service performance evaluation system. The performance evaluation includes the following key components:

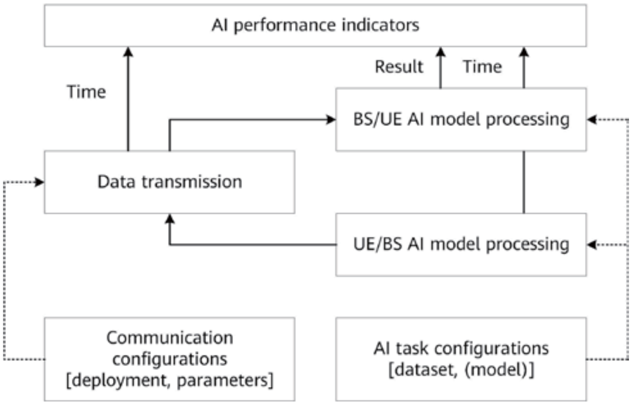

AI performance indicators
Time
Result
Time
BS/UE AI model processing
Data transmission
UE/BS AI model processing
Communication configurations [deployment, parameters]
AI task configurations [dataset, (model)]

Fig. II-6.4.1-1, AI service performance evaluation system

- Resource assumptions: The evaluation should be done in a test environment similar to the definition in communication performance evaluations [6]. Within the test environment, the radio configurations should be provided, including the bandwidth, number of antennas at the UE and BS, and so on.
- AI service procedures: The entire procedures can start from AI model processing at the UE where the intermediate data (model output or model weights) is generated. Then, this data is transmitted from the UE to the BS under the assumed radio configurations. Next, the BS receives the intermediate data and uses the AI model to process it in order to get the service results, which are then used to calculate performance indicators.
- AI service performance calculation: The AI service accuracy and latency are calculated based on the service results, AI model processing time, and transmission time. The AI service accuracy is defined according to the AI task. The AI service latency is the sum of the AI model processing time at the UE and BS and the transmission time of intermediate data.

For AI service density evaluation, AI service density is defined as the number of AI services that meet given AI service accuracy and latency requirements. This can be evaluated through AI service accuracy and latency simulation. For example, we can first set the number of served UEs N to a minimum value, and generate service requests from the UEs. Then, we use the evaluation parameters of the test environment to perform system simulation and collect statistics on the AI service accuracy within the service latency. We can gradually increase N and repeat the simulation until the AI service accuracy falls below requirements, with the value of N to be N_{max} . The AI service density is calculated as $C = N_{max}$/Converge area.

II-6.4.2. Evaluation Example

We use the distributed AI inference service as an example to illustrate the performance evaluation methodology presented earlier. The methodology can also be used for collaborative training and inference services after the procedures are modified according to the corresponding service procedures. Suppose, AI-enabled robots need to perceive the environment through in-factory cameras. The images these robots collect can be further used to achieve real-time high-accuracy AI model inference as in Fig. II-6.4.2-1. The AI inference service consists of three steps: 1) the UE uses the UE-side AI model to process the input data in order to obtain intermediate data; 2) the UE transmits this data to the BS; 3) the BS uses the BS-side AI model to process the received intermediate data and obtain the inference results. The following evaluation methodology can also be applied for this downlink case.

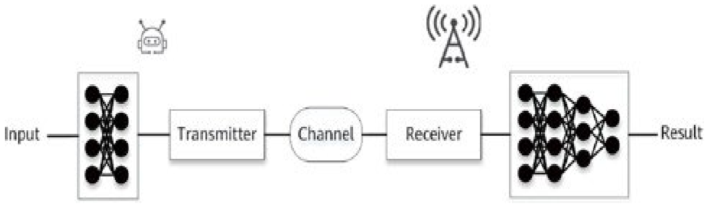


Fig. II-6.4.2-1 Distributed AI inference service example

- Evaluation configurations: The evaluation configurations are defined as follows, with examples given in brackets.
  - Test environment: [Dense Urban]
  - Radio configurations: [same as immersive communication (e. g. , user **experienced data rate: 500 Mbps)]**
  - AI task: [image recognition]
  - AI dataset: [ImageNet-1k validation dataset [7]]
  - AI model: [AlexNet [8], the left part is processed by the UE, and the right part is processed by the BS, as shown in Fig. II-6.4.2-2]
  - AI model processing time: [UE: 0.75 ms; BS: 0.45 ms]
- Evaluation procedures
  - AI service accuracy: AI service accuracy can be evaluated by simulation. The UE processes each input of sample $S_i, i = 1, \cdots, n$ in the data set based on the UE-side AI model, and obtains the intermediate data Z_i.

According to the test environment and transmission configurations, the UE sends the intermediate data and the BS receives it. Taking a classical transmission scheme as an example, the intermediate data is first quantized and represented as bits, which are then encoded and modulated to symbols for wireless transmission. The BS processes the received intermediate data $\tilde{z}_i$ based on the AI model on the BS side, and obtains the inference result $\tilde{Y}_i$ corresponding to each sample. We can then compare or calculate the inference results with the target output or label Y_i of each sample in order to obtain the degree to which the output is the same as the true value $acc = \frac{1}{n}\sum_{i=0}^{n} 1_{\{\tilde{Y}_i==Y_i\}}$, that is, the AI service accuracy.

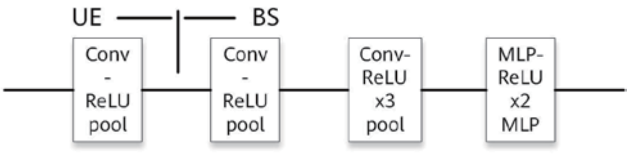


Fig. II-6.4.2-2 AlexNet model deployment example

For the accuracy of the reference case, we can process each sample S_i in the dataset based on the whole AI model in order to obtain the inference result $\widetilde{Y'}_i$. We can then compare the inference result with the label Y_i of each sample to obtain the output of the reference case. The degree to which the output is the same as the true value of reference case is $acc_{ref} = \frac{1}{n}\sum_{i=0}^{n} 1_{\{\widetilde{Y'}_i==Y_i\}}$. The relative AI service accuracy is calculated as acc/acc_{ref}.

– AI service latency: The AI service latency is the sum of the time used for intermediate data transmission, t_{com} and the UE- and BS-side AI model processing time, t_{prc_UE} , t_{prc_BS}. Therefore, the AI service latency is given by $t_{svc} = t_{com} + t_{prc\_UE} + t_{prc\_BS}$. In this example, we use the time calculated as the number of payload bits divided by the data rate as the data transmission time. The number of payload bits is determined by the number of elements in the intermediate data and the number of quantized bits per element.

Table II-6.4.2-1 AI service performance evaluation results

Number of bits per element	2	4	6	8	10	12	16	32
AI service accuracy (%)	0.14	10.35	52.94	56.47	56.53	56.55	56.55	56.56
Relative AI service accuracy (%)	0.24	18.30	93.61	99.84	99.95	99.99	99.99	100
AI service latency (ms)	1.4	1.6	1.8	1.9	2.1	2.3	2.7	4.2

- Evaluation results

The AI service accuracy and latency under different transmission setups (i.e., number of quantized bits per element) are provided in Table II-6.4.2-1. As can be seen from the table, there is a trade-off between AI service latency and AI service accuracy due to the intermediate data transmission. If a minimum AI service accuracy of 56% with maximum AI service latency of 2ms are required (i.e., out target), we need to optimize the transmission configurations of 8 bits per element in this example or improve the transmission technology to meet both requirements.

II-6.5. Conclusion

This paper illustrated the motivations, typical AI services, and performance requirements of the "AI and Communication" usage scenario — a new scenario defined in IMT-2030 for 6G. To provide guidelines for the system design and better support AI services, this proposed new performance indicators that integrate AI and communication capabilities and resources in the network, from both the user experience and network capacity perspectives. It also provided the corresponding evaluation methodology with a detailed example. For further detail of the contexts, the original paper can be seen in [9].

REFERENCE

[1] ITU-R, Recommendation ITU-R M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond," Nov. 2023.

[2] 3GPP TR 22.874, "Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," Release 18, 2021.

[3] IMT-2030 (6G) Promotion Group, "White paper on typical usage scenarios and key capabilities in 6G," July 2022.

[4] Hexa-X Deliverable D1. 3, "Targets and requirements for 6G – Initial E2E architecture," Feb. 2022.

[5] https://mlcommons.org/en/, accessed on Aug. 10, 2022.

[6] ITU-R M.2412-0, "Guidelines for evaluation of radio interface technologies for IMT-2020," 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," IEEE CVPR, 2009.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS, 2012.

[9] Gongzheng Zhang, Jian Wang, and Rong Li, et.al., “Performance Requirements and Evaluation Methodology for AI and Communication in 6G”, Communication of HUAEWI RESEARCH, November 2024, pp64-72.

II-7. Study on AP Clustering with Deep Reinforcement Learning for Cell-Free Massive MIMO

Yu Tsukamoto, Akio Ikami, Takahide Murakami,
Amr Amrallah, Hiroyuki Shinbo, and Yoshiaki Amano
KDDI Research, Inc.

Abstract— **Cell-free massive MIMO (CF-mMIMO) is a promising approach for future mobile networks, utilizing centralized MIMO processing for densely distributed access points (APs). In CF-mMIMO, to reduce the computational load for signal processing while meeting throughput demands, user equipment (UEs) are served by APs selected as an AP cluster. A significant challenge is AP clustering for each UE, particularly in dynamic environments with moving UEs. One approach for optimizing the AP cluster involves AI/ML. This paper provides an overview of AP clustering method using deep reinforcement learning (DRL) and numerical simulation results.**

II-7.1. Introduction

A larger-scale distributed MIMO architecture, i.e., cell-free massive MIMO (CF-mMIMO), has gained attention as a technology capable of providing high radio quality throughout an entire area [1]-[3]. In CF-mMIMO, a central processing unit (CPU) performs multi-user MIMO processing on radio signals from distributed APs. By coordinating signal processing among distributed APs, interference at the cell edges can be significantly reduced, ensuring high radio quality throughout the service area. APs with weak channel strength do not significantly enhance the radio quality for UEs. In AP clustering, APs that enhance radio quality are selected to form AP clusters for each UE, and the UE is served only by these selected APs [4]-[6]. This reduces the signal processing load since the number of channels involved in the MIMO calculation is limited.

To reduce the signal processing load while ensuring the required radio quality for each UE, selecting AP clusters appropriately, on the basis of the movement of the UE, is crucial. Recently, the use of DRL for AP clustering has been explored in [7]-[15]. However, scalability continues to be a challenge. In large-scale environments with numerous APs and UEs, the computational load of DRL becomes significant. In this paper, we provide the overview of scalable AP clustering method with DRL and present the AP clustering performance in terms of throughput requirement satisfaction and computational load, including signal processing, as well as training and inference in DRL.

II-7.2. AP Clustering Problem for CF-mMIMO

We have been researching user-centric radio access network (RAN) architecture to ensure consistent radio quality throughout the network area using CF-mMIMO [16]. The user-centric RAN concept involves creating a logical network for each user on a physical infrastructure through network virtualization, as illustrated in Fig. II-7.2-1. Virtualized CPUs (vCPUs) are deployed for each user on the servers, with APs linked to these servers via mobile fronthaul. The vCPU executes multi-user MIMO by using radio signals to and from the APs within each user's AP cluster. The radio quality, measured as the signal-to-interference-plus-noise ratio (SINR), is influenced by the channels of the selected APs in the AP cluster and other spatially multiplexed UEs.

When DRL is applied to AP clustering in large-scale environments, the computational load of DRL increases because of two main factors. The first factor is the increase in model size. When DRL is used to select the optimal combination of APs and UEs, the size of the action space, that is, the number of candidate action combinations, expands exponentially with increasing number of APs and UEs. Moreover, as information about the entire area is required as an input for the model as states. These result in a larger neural network (NN) size in large-scale environments. The second factor is the increase in inference frequency. As the wireless environment changes with UE movement, it is necessary to dynamically select AP clusters to maintain the radio quality of UEs. Selecting AP clusters for all UEs at short intervals increases the overall inference frequency, thereby escalating the computational load of inference across the system.

Our goal is to ensure the required radio quality for each UE with minimal computational load even in large-scale environments. To facilitate scalability, an AP clustering approach is necessary to suppress the overall computational load, including training and inference in DRL, as well as signal processing.

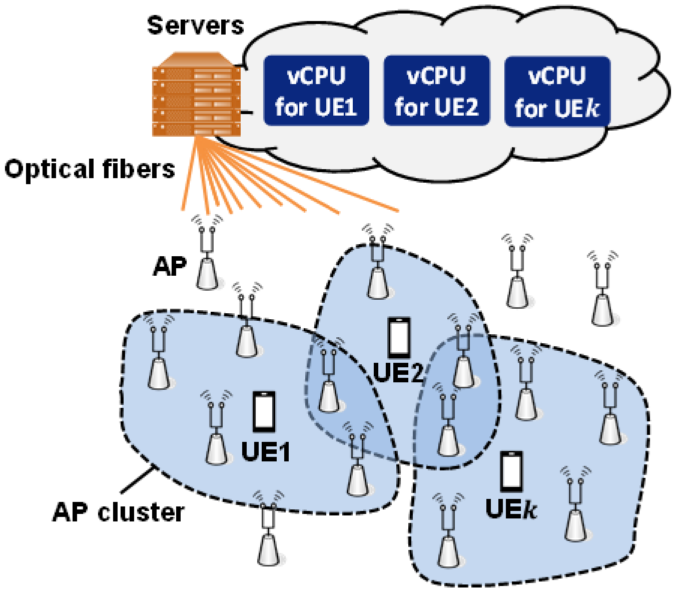


Fig. II-7.2-1: User-centric RAN architecture with CF-mMIMO

II-7.3. Scalable AP Clustering with Distributed DRL

II-7.3.1. AP Clustering Architecture

We propose an AP clustering method with distributed DRL, including the following two components. The first component involves distributing per-user models. The learning model is designed to determine the increment or decrement of the AP cluster size for a single UE. Since APs with higher channel strength enhance radio quality, the combination of APs is then selected in descending order of the reference signal received power (RSRP) up to the determined AP cluster size. The model dynamically adjusts the AP cluster size for each UE. The proposed per-user model maintains a constant size, independent of the number of UEs or APs, preventing any increase in model size even in large environments. To achieve real-time AP clustering with high learning efficiency, the per-user model is distributed and processed in parallel. We use Ape-X [17], a distributed learning method for DRL. Fig. II-7.3.1-1 illustrates the proposed AP clustering architecture. In Ape-X, agents are divided into actors and learners. Multiple actors observe the state of the environment and determine actions in parallel via a common learning model provided by the learner. The learner performs training and updates the model from experiences generated in parallel by multiple actors.

The second component involves assigning UEs to the actors. If an actor is launched for each UE and performs inferences with a short cycle, the overall inference frequency increases, especially in large-scale environments with many UEs. For fast-moving UEs, a short AP cluster update interval is necessary to maintain radio quality. However, for slow-moving UEs, longer intervals do not significantly impact radio quality. Therefore, as shown in Fig. II-7.3.1-1, we introduce an actor allocator (AA) to assign multiple slow-moving UEs to the same actor. Since the learning model operates on a per-user basis, the actor with multiple UEs assigned performs inference sequentially. This increases the AP cluster update interval for the UEs. To avoid throughput degradation, the maximum update interval that does not degrade the radio quality is defined as the threshold interval. UEs are assigned to actors under the constraint that their AP cluster update interval does not exceed the threshold interval, minimizing the number of actors. This approach reduces the computational load of inference by relaxing the inference frequency.

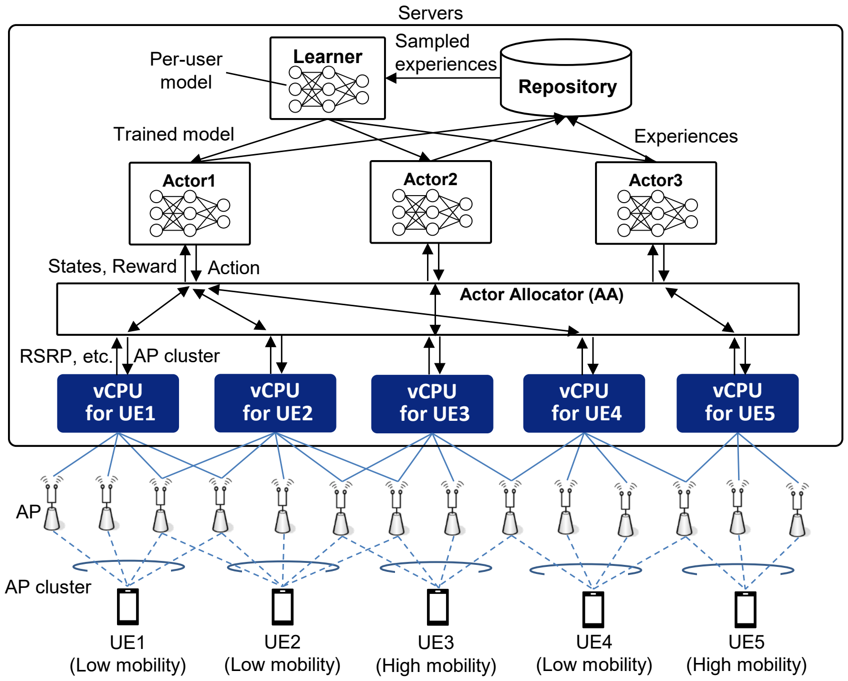


Fig. II-7.3.1-1: AP clustering architecture with distributed DRL

II-7.3.2. MDP Model

The design goal of the Markov decision process (MDP) model is to control the AP cluster size for a single UE to meet throughput requirements with the minimum AP cluster size. The action, reward, and state in the MDP model are defined as follows:

1) Action

The action specifies the increment or decrement in the AP cluster size. The action for UE k is defined as $a_k = \delta_k \in \{-e, -e+1, \dots, 0, \dots, e-1, e\}$. Here, δ_k represents the change in the AP cluster size for UE k from the previous time step. e denotes the maximum change in the AP cluster size in one time step. The AP cluster size $|\mathcal{M}_k|$ is determined as $|\mathcal{M}_k| = \delta_k + |\mathcal{M}_k|^{\mathrm{pre}}$, where $|\mathcal{M}_k|^{\mathrm{pre}}$ represents the AP cluster size at the previous time step. The size of the action space $|A_k|$ is $2e+1$.

2) Reward

We use the following reward r_k, which consists of two factors: throughput satisfaction and the AP cluster size.

$$r_k = q_k \times m_k$$

where q_k and m_k are defined as:

$$q_k = \begin{cases} 1 & g_k \geq \tilde{g}_k, \\ 0 & \text{otherwise.} \end{cases}, \quad m_k = \left(1 - \frac{|\mathcal{M}_k|}{L}\right)^3$$

q_k indicates throughput satisfaction, where $\tilde{g}_k$ is the preset throughput requirement for UE k. If the throughput g_k does not meet this requirement, the reward is 0. m_k indicates the AP cluster size factor. The computational load for signal processing is proportional to the cube of the AP cluster size $|\mathcal{M}_k|$. m_k decreases in proportion to the cube of the AP cluster size. L is the number of APs in the area. The reward is high when the throughput requirements are met with the minimum AP cluster size for UE k.

3) State

The state for UE k is defined as $S_k = \left[|\mathcal{M}_k|^{\mathrm{pre}}, \tilde{g}_k, U_k, U_k^{\mathrm{pre}}, j_k\right]$. The previous AP cluster size $|\mathcal{M}_k|^{\mathrm{pre}}$ is needed to determine the change in the AP cluster size from the previous time step. The throughput requirement $\tilde{g}_k$ helps ascertain the required radio quality for the UE. $U_k = \left[u_{k,1}, u_{k,2}, \dots, u_{k,b}, \dots, u_{k,B}\right]$, where $u_{k,b}$ is the RSRP from the b-th highest AP. U_k^{pre} represents U_k at the previous time step and helps to learn changes in the channel state due to UE mobility. To account for the impact of other UEs around UE k, we employ j_k as the count of overlapping APs in the AP cluster of UE k and other UEs.

II-7.4. Simulation Evaluation

II-7.4.1. Simulation Conditions

The main parameters for the numerical simulation are summarized in Table II-7.4.1-1. We use a 1 km^2 urban structure with 400 APs around Shibuya Station in Tokyo and employ channel data based on ray tracing. 100 UEs with different throughput requirements and velocities move randomly. The throughput requirements and velocities of each UE are randomly set from {50, 100, 150} Mbps and {0, 4, 30, 60} km/h, respectively.

Table II-7.4.1-1: Simulation parameters

Parameters	Values
RAN environment parameters	
Simulation area (max)	1 km×1 km at Shibuya in Tokyo
Number of deployed APs, L	400
Number of UEs, K	100
Number of antennas in AP, N	1
Frequency	3.5 GHz
System bandwidth	100 MHz
UE transmission power	20 dBm
Large-scale fading	Ray tracing
Small-scale fading	Rayleigh fading
Noise figure	7 dB
Number of pilot sequences, τ_p	24
UE movement speed, ψ_k	{0, 4, 30, 60} km/h
User traffic	Full buffer
Throughput requirements, $\tilde{g}_k$	{50, 100, 150} Mbps
Time step length	50 msec
GA parameters	
Population size	50
Number of generations	200

Mutation rate	0.2
DRL parameters	
Target network update intervals	2500
Network parameters copy intervals	500
Training batch size	512
Discount factor	0.5
Learning rate	0.00025/4
Episode length	1000 time steps (50 seconds)
Number of training episodes	100
Number of test episodes	5
Number of RSRP in state, B	20
Variation range in action, e	2

II-7.4.2. Evaluated Methods

In the simulation evaluation, we compare the following methods:

- Static approach (SA): The AP cluster size for each UE is predetermined. To satisfy throughput requirements with 90% probability, we set AP cluster sizes of 5, 7, and 13 for UEs with throughput requirements of 50 Mbps, 100 Mbps, and 150 Mbps, respectively.
- Closed-loop control (CLC): If the throughput does not meet the requirements, the AP cluster size is increased. Conversely, if the throughput meets the requirements, the size is decreased.
- Genetic algorithm (GA): We define the combination of the actions for each UE as an individual. The objective function is defined as the summation of rewards for all UEs. The parameters for GA are shown in Table II-7.4.2-1.
- Distributed DRL (D-DRL): To validate the effectiveness of the AA, we introduce D-DRL without the AA. We adopt the architecture of D-DRL via the per-user model described in Section I-1.3. .
- Distributed DRL With Actor Allocator (D-DRL with AA): This is a proposed method for applying AA to D-DRL.

II-7.4.3. Simulation Results

Fig. II-7.4.3-1 shows the average throughput satisfaction rate, indicating the ratio of UEs meeting throughput requirements among all UEs. It is defined as $\sum_{k \in K} q_k / K$. SA, D-DRL, D-DRL, D-DRL with AA, and GA maintain a throughput satisfaction rate of around 90%. In SA, a fixed AP cluster size is set to satisfy throughput requirements with a 90% probability. In D-DRL and D-DRL with AA, the satisfaction rate is kept at the same level as that of SA, and approaches that of GA which obtain near-optimal solutions. CLC based solely on throughput feedback makes it difficult to consistently satisfy the throughput requirements.

Fig. II-7.4.3-2 presents the total computational load of signal processing, inference, training and GA. D-DRL and D-DRL with AA suppress the signal processing load compared with SA and CLC by selecting the minimal AP cluster size for each UE. In D-DRL, actors are launched for each UE in parallel. This increases the inference load proportionally to the number of UEs. For D-DRL with AA, the inference load is lower than that of D-DRL because AA minimizes the number of launched actors and reduces the inference frequency. GA needs substantial computational resources for real-time AP clustering. The total computational load for D-DRL with AA is reduced by 29% compared with that of SA. The proposed method demonstrates AP clustering to facilitate scalability in large-scale environments.

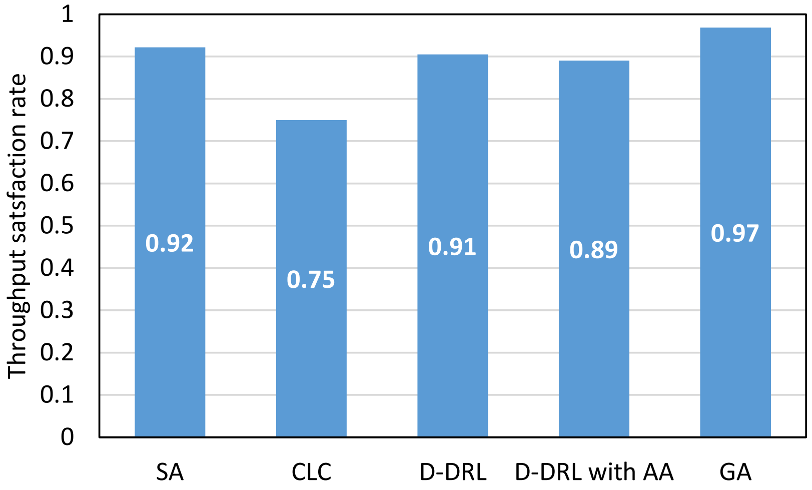


Fig. II-7.4.3-1: Throughput satisfaction rate

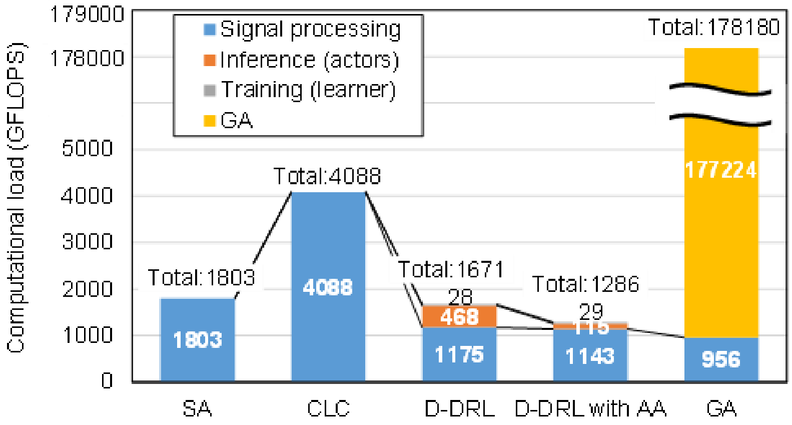


Fig. II-7.4.3-2: Total computational load

II-7.5. Conclusion

In this paper, we introduced an AP clustering method using D-DRL to address the scalability issue. By employing the per-user model and distributed processing, we demonstrated that learning performance remains high with a small-sized model. Furthermore, by assigning UEs to actors on the basis of their movement speed, the inference frequency can be reduced. The overall computational load, including DRL and signal processing, was reduced by 29% compared with that of SA with a fixed AP cluster size. The proposed method achieves AP clustering that satisfies the throughput requirements with minimal computational load, even in large-scale environments.

Acknowledgement

These research results were obtained from the research (JPJ012368C00401) commissioned by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCE

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO: Uniformly great service for everyone,” in 2015 IEEE 16th Intern, Workshop on Signal, Process, Advances in Wirel, Commun,, Stockholm, Sweden, June 2015, pp.201-205.

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-free massive MIMO versus small cells,” IEEE Trans.Wireless Commun., vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[3] E. Björnson and L. Sanguinetti, “Making cell-free massive MIMO competitive with MMSE processing and centralized implementation,” IEEE Trans. Wireless Commun., vol. 19, no. 1, pp. 77–90, Jan. 2020.

[4] S. Buzzi and C. D’Andrea, “Cell-Free Massive MIMO: User-Centric Approach,” in IEEE Wirel. Commun. Lett., vol. 6, no. 6, pp. 706-709, Dec. 2017.

[5] G. Interdonato, P. Frenger, and E. G. Larsson, “Scalability aspects of cell-free massive MIMO,” in IEEE Int. Conf. on Commun. (ICC), pp. 1–6, May 2019.

[6] E. Bjornson and L. Sanguinetti, “Scalable Cell-Free Massive MIMO Systems,” IEEE Trans. Commun., vol. 68, no. 7, pp. 4247-4261, Jul. 2020.

[7] X. Chai, H. Gao, J. Sun, X. Su, T. Lv and J. Zeng, "Reinforcement Learning Based Antenna Selection in User-Centric Massive MIMO," in 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-6.

[8] R. Y. Chang, S. -F. Han and F. -T. Chien, "Reinforcement Learning-Based Joint Cooperation Clustering and Content Caching in Cell-Free Massive MIMO

Networks," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Norman, OK, USA, 2021, pp. 1-7.

[9] Y. Al-Eryani, M. Akrout and E. Hossain, "Multiple Access in Cell-Free Networks: Outage Performance, Dynamic Clustering, and Deep Reinforcement Learning-Based Design," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 4, pp. 1028-1042, April 2021.

[10] C. F. Mendoza, S. Schwarz and M. Rupp, "User-Centric Clustering in Cell-Free MIMO Networks using Deep Reinforcement Learning," ICC 2023 - IEEE International Conference on Communications, Rome, Italy, 2023, pp. 1036-1041.

[11] N. Ghiasi, S. Mashhadi, S. Farahmand, S. M. Razavizadeh and I. Lee, "Energy Efficient AP Selection for Cell-Free Massive MIMO Systems: Deep Reinforcement Learning Approach," in IEEE Transactions on Green Communications and Networking, vol. 7, no. 1, pp. 29-41, March 2023.

[12] Z. Gao, Q. Zhang, J. Liu, Z. Du and Y. Li, "DRL-Based AP Selection in Downlink Cell-Free Massive MIMO Network With Pilot Contamination," in IEEE Communications Letters, vol. 28, no. 6, pp. 1432-1436, June 2024.

[13] T. Fangqing, Q. Deng, and Q. Liu, "Energy-efficient access point clustering and power allocation in cell-free massive MIMO networks: a hierarchical deep reinforcement learning approach." EURASIP Journal on Advances in Signal Processing 2024, no. 18.

[14] B. Banerjee, R. C. Elliott, W. A. Krzymieñ and M. Medra, "Access Point Clustering in Cell-Free Massive MIMO Using Conventional and Federated Multi-Agent Reinforcement Learning," in IEEE Transactions on Machine Learning in Communications and Networking, vol. 1, pp. 107-123, 2023.

[15] Z. Liu, J. Zhang, Z. Liu, D. W. K. Ng and B. Ai, "Joint Cooperative Clustering and Power Control for Energy-Efficient Cell-Free XL-MIMO with Multi-Agent Reinforcement Learning," in IEEE Transactions on Communications, 2024.RP-243244, "Revised WID: Artificial Intelligence (AI) / Machine Learning (ML) for NR Air Interface," Qualcomm, December 2024.

[16] K. Yamazaki, T. Ohseki, Y. Amano, H. Shinbo, T. Murakami and Y. Kishi, "Proposal for a User-Centric Ran Architecture Towards Beyond 5G," in 2021 ITU Kaleidoscope: Connecting Physical and Virtual Worlds (ITU K), Geneva, Switzerland, 2021, pp. 1-7.

[17] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, and D. Silver, "Distributed prioritized experience replay," Apr. 2018, arXiv:1803.00933.

II-8. Cross-layer Access Control Techniques using AI

Hiromichi TOMEBA and Osamu Nakamura
Sharp Corporation

***Abstract*— The demand for large-capacity, latency sensitive applications such as ultra-high-definition video transmission is increasing in wireless communication systems. In high-demand applications such as ultra-high-definition video transmission, there is a problem that increasing the capacity of wireless communication does not necessarily lead directly to the realization of such applications. Therefore, to improve the number of applications that can be realized, we have studied the cross-layer access control techniques, which improves video throughput based on whether the requirements for ultra-high-definition video are satisfied, rather than conventionally called throughput, which is calculated from correctly received bits. In this contribution, we show the computer simulation and indoor filed trial results of the radio resource allocation techniques using the video throughput.**

II-8.1. Introduction

Transmission of ultra-high-definition images and videos is increasingly being adopted across in a variety of applications, including industrial fields such as image inspection, healthcare fields. This trend is expected to accelerate in coming years. In the beyond 5G era, it is necessary to facilitate the transmission of large-capacity and low-latency traffic generated by these applications through wireless communication networks. In particular, since ultra-high-definition video is expected to be predominantly used indoors, both the cellular networks and private local area networks (LANs) will play an important roles in supporting these high-demand applications.

On the other hand, user experience is important for application accommodated. For high-demand applications such as ultra-high-definition video transmission, increasing the capacity of wireless communication does not necessarily lead directly to the effective support of these application [1]. Therefore, we have investigated a cross-layer collaborative access control technology that improves the video throughput with the requirements of video transmission considered [2]-[4]. These technologies demonstrates significant improvements in the number of applications accommodated compared to conventional throughput-based methods that consider the number of successfully transmitted bits, such as proportional fairness [5].

As mentioned earlier, considering application requirements contributes to increasing the number of applications that can be accommodated, but the requirements of each application are diverse. For example, even in ultra-high-definition video transmission, multiple requirements such as remaining buffer capacity and allowable delay time are required. Naturally, since there are multiple users, the number of combinations of factors to be considered is enormous. Therefore, we have considered resource allocation using deep reinforcement learning, considering the use of machine learning.

This contribution explain the concept of application layer throughput and show the simulation and experimental results.

II-8.2. System Model

Conventionally, the throughput is generally used as a performance metric of wireless communication, which is calculated from the number of bits included in successfully received transmission packets (hereafter referred to as bit throughput). However, user experience is necessary for the accommodation of video applications, and it is required to satisfy the requirements of the applications. Therefore, it is important to use video throughput which is the throughput calculated from the received video packets satisfying the requirements as a performance metric [6]. Fig. II-8.2-1 shows an overview of video throughput. The requirements for video throughput are that periodically generated video packets must be correctly received within the allowable delay time. If a video packet cannot be transmitted within the allowable delay time, some of the correctly received transmission packets are counted as bit throughput, but they are not counted as video throughput. If the requirement of receiving a video packet within the allowable delay time is satisfied, the number of bits included in the video packet is counted as video throughput.

In general, video data is divided into multiple video packets and transmitted. Conventional bit throughput is calculated from the number of bits included in successfully received transmission packets. Here, the bit throughput of the u-th user is $R_u^b(T)$, and the video throughput is $R_u^v(T)$.

$$\begin{cases} R_u^b = \dfrac{1}{T}\displaystyle\sum_{t=0}^{T-1} B_u^b(t)\,, \\ R_u^v = \dfrac{1}{T}\displaystyle\sum_{t=0}^{T-1} B_u^v(t) \end{cases} \tag{1}$$

where T is the observation time, $B_u^b(t)$ is the number of bits correctly received by the u-th user from time t-1 to t, and $B_u^v(t)$ is the video packet size correctly received by the u-th user from time t-1 to t.

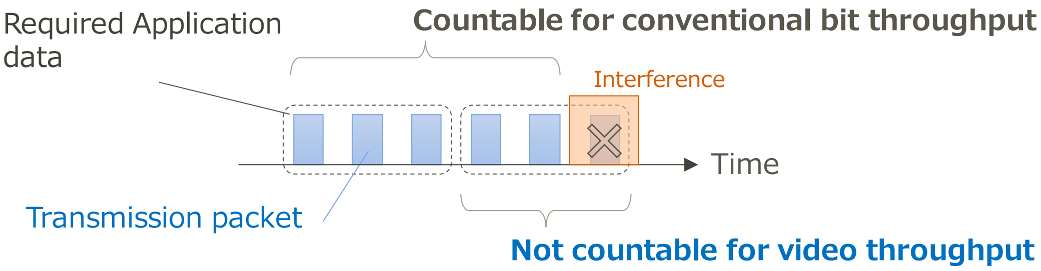

Required Application data
Countable for conventional bit throughput
Interference
Time
Transmission packet
Not countable for video throughput

Fig. II-8.2-1. Relation between the application and physical layer throughput.

In order to improve video throughput, it is necessary to appropriately allocate wireless resources by considering the infinite number of user conditions, such as remaining packet size, allowable delay time, allocatable wireless resource candidates, and their quality. Using the video throughput shown above as a metric, we have been studying the allocation of frequency bands in which each user exchanges frames and the allocation of available radio resources within the selected frequency band.

Regarding radio resource allocation, two methods are considered: wideband (WB) resource allocation, which allocates the entire available wireless resource band to one user, and SB resource allocation, which divides the entire band into multiple subbands (SB) and allocates each SB to a user. In wideband transmission where frequency-selective fading cannot be ignored, SB resource allocation is suitable for increasing the capacity of the system because it can provide user diversity effects. However, as the number of combinations increases relative to WB resource allocation, the complexity of scheduling also increases.

For the SB resource allocation, we have studied the radio resource allocation techniques with the deep reinforcement learning (RL) using deep neural network (DNN) considering the user situation including the achievable the video throughput. In order to consider the video throughput, we set the reward r for RL expressed as

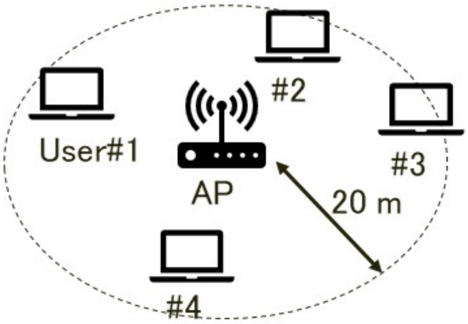
$$r = \sum_{u=1}^{N_u} r_u \, , r_u = \begin{cases} 1, & R_u^v/\hat{R}^v \geq \alpha \\ 0, & R_u^v/\hat{R}^v < \alpha \end{cases} , (2)$$

where N_u, $\hat{R}^v$, and α are denote the video throughput of u-th user, average video rate, and threshold for the ratio of the current video throughput to the average video rate, respectively. The reward r leads the number of applications accommodated increase.

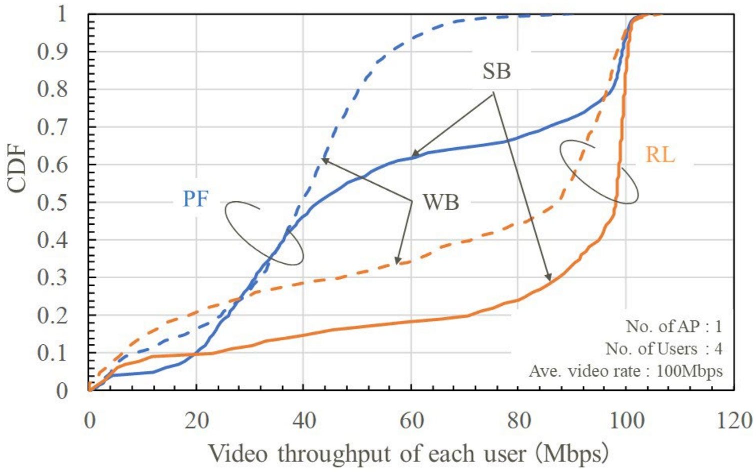
Figure II-8.2-2(a) represents the assumed system model. In this simulation, we assume the downlink transmission using infrastructure mode where the single access point (AP) with 4 users. Simulation conditions are summarized below. The carrier frequency and the bandwidth are assumed as 5 GHz and 80 MHz. The candidates of the SB size are 20, 40 and 80 MHz. The transmit power is 20 dBm. The antenna gain is 0 and -2 dBi for AP and each user. The standard deviation and noise figure are 5dB and 7 dB, respectively. The average video rate is 100 Mbps, where the video packet are periodically generated with 10 ms periodicity. The video traffic is generated using the wireless display model [8]. Other conditions follow the evaluation scenario of IEEE 802.11ax [7].

In the RL, the assumed learning algorithm is DQN. The number of epochs is set to 200, and the learning model is updated every epoch. Regarding STA placement, to acquire generalization performance, STAs are randomly placed between epochs during the learning process. In each epoch, 500 steps are executed. The DNN for actor consists of an input layer, three hidden layers, and an output layer. Each hidden layer has 32 nodes. The reward parameter α of (2) is set to 0.25 based on simulations. Regarding STA placement, to acquire generalization performance, STAs are randomly placed between episodes during the learning process.

Figure II-8.2-2(b) shows the cumulative distribution function (CDF) of video throughput of each user. For comparison, the conventional proportional fairness (PF), which consider the bit throughput, is also shown in Fig. II-8.2-2(b). It is noted that the maximum video throughput is not meet the average video rate since the video packet size is randomly generated. It is shown that the SB allocation of the video throughput is better than that of the WB allocation irrespective of the RL and PF. It is also seen from Fig. II-8.2-2(b), the RL can provide a better video throughput performance than the PF. From the simulation results, in terms of the number of applications accommodated (calculated as the number of users with an average video rate of 90% or more), RL using SB allocation can provide approximately 2.4 times better performance compared to PF using SB allocation.



(a) Simulation model



(b) Video throughput of each user

Fig. II-8.2-2. Simulation results.

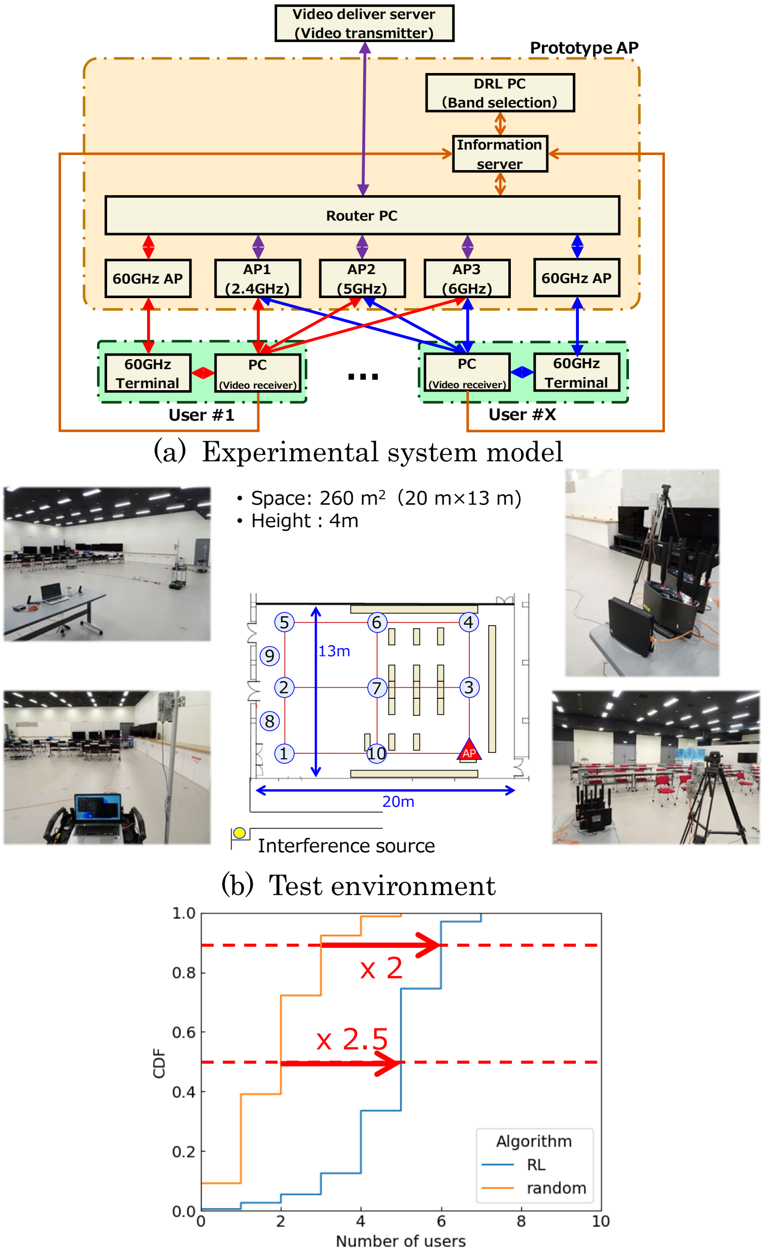
II-8.3. Experimental Result

We are conducting indoor propagation environment tests using a prototype that utilizes wireless LAN equipment for cross-layer access control technology based on deep reinforcement learning, which has demonstrated improvement effects through computational simulations [6]. The outline of the demonstration system is shown in Figure II-8.3-1(a). To simulate an AP capable of controlling multiple frequency bands, including the millimeter-wave band at 60 GHz, we connected multiple APs configured for each frequency band to a control PC (router PC), which was played as the prototype AP. The outline of the testing environment is illustrated in Figure II-8.3-1(b). Assuming a medium-sized conference room, the prototype AP was placed in a corner of the environment, and users equipped with video receivers were positioned at points labeled 1-th to 10-th in Figure II-8.3-1(b). Each user supports 2.4/5/6 GHz, with users at points 1-th and 2-th specifically supporting 60 GHz as well.

In each frequency band, excluding the 60 GHz band, the available bandwidth is set to 20 MHz, while the 60 GHz band is set to 2.16 GHz. The average video rate is assumed as 60 Mbps. Additionally, to simulate interference from other systems, a pair of access points (APs) and users as interference sources are separately deployed, generating an average interference traffic of 100 Mbps in the 6 GHz band.

The decoded results of the video packets transmitted to each user are fed back to the prototype AP as video throughput. The deep reinforcement learning PC determines the frequency band allocated to each user based on the feedback, allowing for the evaluation of performance characteristics. For comparison, a scenario is also tested where the frequency bands are selected randomly without employing reinforcement learning based on video throughput.

Figure II-8.3-1(c) presents the cumulative distribution function (CDF) of the number of users that can be accommodated, where the number of accommodated users is defined as those satisfying the average video rate of 60 Mbps. As illustrated in Figure II-8.3-1(b), when utilizing deep reinforcement learning, an improvement of approximately 2.5 times at the CDF 50% value and approximately 2 times at the CDF 90% value is observed compared to the random selection scenario. It can be confirmed that deep reinforcement learning effectively adapts to the propagation environment and the achievable video throughput.



(a) Experimental system model

(b) Test environment

(c) Experimental result

Fig. II-8.3-1. Simulation results.

II-8.4. Conclusion

This contribution shows the concept of application layer throughput for improving the number of accommodated users. We show the simulation and experimental results and the application layer throughput can improve the number of accommodated users.

Acknowledgement

This work includes the results of the "Research and Development for the Expansion of Radio Wave Resources (JPJ000254)" of the Ministry of Internal Affairs and Communications.)

REFERENCE

[1] E.-K. Hong, et al., "6G R&D vision: Requirements and candidate technologies," Journal of Communications and Networks, vol. 24, no. 2, pp. 232-245, Apr. 2022.

[2] O. Nakamura, et.al., "Evaluation of Radio Frequency Band Selection using Deep Reinforcement Learning Based on Video Transmission Requirements under Interference Environment," IEICE Technical report, vol. 121, no. 391, RCS2021-275, pp. 127-132, March 2022

[3] R. Yamada, H. Tomeba, T. Sato, O. Nakamura, and Y. Hamaguchi, "A Radio Resource Allocation Technique using Requirements for Video Transmission," IEICE Communications Express (ComEX), https://doi.org/10.1587/comex.2022COL0018.

[4] R. Yamada, H. Tomeba, T. Sato, O. Nakamura, and Y. Hamaguchi, "Uplink Resource Allocation for Video Transmission in Wireless LAN System," IEEE 8th World Forum on Internet of Things, Oct. 2022.

[5] R. Yamada, et.al., "Study on sub band resource allocation considering the video transmission requirement using the reinforcement learning,", IEICE Society conference 2024, B-5B-02, Sep. 2024

[6] H. Tomeba, "Cross-layer access control techniques for next generation wireless network," IEICE Technical report, RCS2024-185, Dec. 2024.

[7] S. Merlin, et al, "TGax Simulation Scenarios," IEEE 802.11-14-0980, July 2015.

[8] G. Li, "A Few Corrections to Video Traffic Model," IEEE 802.11-14-1440, Nov. 2014.

II-9. AI-based Application-aware RAN Optimization

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

***Abstract*— It has become increasingly important for industries to promote digital transformation by utilizing 5G/6G, Internet of Things (IoT), and Artificial Intelligence (AI) to realize a highly productive and prosperous society. In addition to conventional policies of improving the average Quality of Service (QoS) at each mobile coverage area, there is an increasing need to strengthen policies that precisely adhere to QoS requirements per User Equipment (UE) and in real-time to enable the stable use of applications at high-performance levels, e.g., work speed or productivity. The Open Radio Access Network (Open RAN), specifically standardized by the O-RAN Alliance (O-RAN), offers significant potential to enable flexible resource management to address diverse QoS requirements. This article introduces an application-aware RAN optimization method that can support such policies based on O-RAN architecture.**

II-9.1. 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 5G/6G, IoT, and AI to realize a highly productive and prosperous society [1][2][3][4]. In digital transformation, many use cases require mobility and ease of equipment installation, making reliable wireless communication essential. To enable the stable use of applications at high-performance levels, e.g., work speed or productivity, often results in stricter QoS requirements. Thus, in addition to policies aimed at improving the average QoS at each mobile coverage area, there is an increasing need to strengthen policies that precisely adhere to QoS requirements for each UE in real-time.

Application developers often design applications based on current wireless communication standards, whereas innovative developers focus on application goals first and then address communication issues through trial and error. Customized communication infrastructures are often required for specific applications, which are not scalable for widespread 5G/6G adoption. To this end, the RAN must be autonomously and adaptively controlled based on the application, network, and site conditions. Emerging trends like Open RAN, specifically standardized by the O-RAN [5][6], offer significant potential to enable flexible resource management to address diverse QoS requirements.

This article introduces an application-aware RAN optimization method based on the O-RAN architecture to support such policies [7][8].

II-9.2. AI Native Open RAN

Open RAN is the concept of disaggregating functions within the RAN, enabling the various hardware and software functions that make up the RAN to be provided in a multi-vendor, interoperable environment. Open RAN is an ongoing shift in mobile network architectures for operators to introduce non-proprietary subcomponents from various vendors that adhere to a set of industry-wide standards that telecom suppliers can follow when producing related equipment. The O-RAN is a worldwide community of operators and vendors with a mission to reshape RAN to be more open, virtualized, and fully interoperable. One of the key advantages of Open RAN is the introduction of greater automation and intelligence into networks. The use of AI-driven capabilities and virtualized computing and distribution functions will lead to a significant reduction in hardware-dependent systems. The introduction of other functions, such as "rApps" running on non-real-time and "xApps" running on near-real-time RAN Intelligence Controllers (RIC) platforms, will help operators intelligently monitor and manage their networks.

Many operators and vendors have provided their visions for 6G. Most of these visions emphasize two critical points: automation and the need for 6G to be “AI native”. Given the higher speeds and lower latencies involved, most anticipate even more automation and intelligence in 6G. The plans for solutions that 6G will support are already being developed today in Open RAN, which is expected to serve as a critical architectural foundation for 6G, much like virtualization is a foundational element for 5G RAN today.

Delivering new 5G/6G solutions to new markets requires collaboration with industry vertical vendors and other specialist vendors, which in turn requires the open architecture and collaborative models that Open RAN provides. A good example is that mobile operators struggled in the past to provide bespoke solutions that could meet the specific needs of individual enterprises. With 5G/6G and Open RAN capabilities, it is now possible to deliver services that are tailored to the individual enterprise’s needs and create new business opportunities for operators.

II-9.3. Application-aware RAN Optimization

The industrial use case is considered to ensure the uninterrupted transport of materials at the factory/warehouse floor utilizing Autonomous Mobile Robots (AMRs). Regarding latency, availability, and determinism, communication services for remote-control applications must fulfill stringent requirements. In these applications, cyclic two-way communication is essential for monitoring robot status and sending control instructions. If the latency exceeds a certain threshold, the system will safely stop to ensure safety. While fail-safes are essential for maintaining safety, frequent occurrences of these fail-safes can lead to decreased facility utilization and productivity.

The application-aware RAN optimization method utilizes AI to analyze the communication requirements and radio quality fluctuations for individual UEs, including robots and vehicles. Based on this analysis, the AI dynamically adjusts RAN parameters for each UE to optimize performance. This AI learns from past operational records of robots and vehicles to optimally control the RAN parameters. It adjusts RAN parameters such as the target block error rate, the allocation ratio of physical resource blocks, and the allowable additional delay while predicting the likelihood of exceeding communication latency requirements. In typical 5G networks, RAN parameters are fixed and configured for the entire network. However, the proposed method dynamically adjusts them per-UE basis to improve application productivity. The architecture is shown in Fig. II-9.3-1. The proposed method can perform the following tasks for each UE basis in near-real-time: 1) estimating application QoS requirements based on information supplied by the external application server, 2) predicting fluctuations in wireless quality using radio quality data from the central unit (CU) and distributed unit (DU), and 3) proactively optimizing CU and DU parameters. While running machine learning, the system ensures that accuracy is uncompromised. If a risk is detected, it switches to a stable logic-based engine. This technology guarantees stability in RAN control by switching engines.

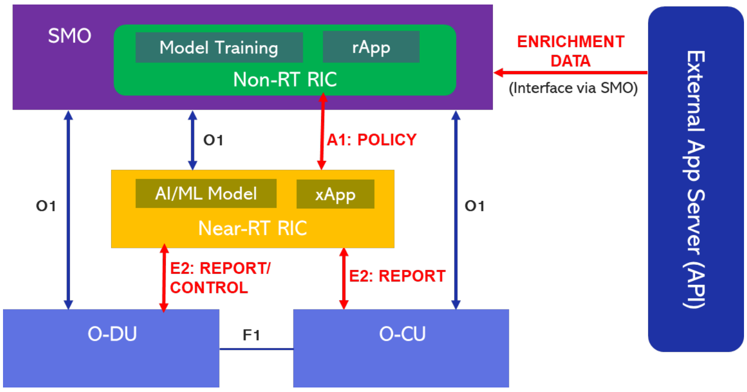


Fig. II-9.3-1 Architecture

II-9.4. Evaluation

The simulation was conducted in which a server remotely controlled mobile robots over a 5G network indoors. We developed and utilized a precise simulator consisting of mobility, radio propagation, and network simulators. The timeliness and availability of communication services were evaluated in the context of mean time between failures in mobile robot operations, one of the critical key performance indicators defined by 3GPP for this type of traffic. The proposed method optimizes RAN parameters, including the target block error rate, the allocation ratio of physical resource blocks, and the allowable additional delay each component can endure on a per-UE basis in near real-time. The proposed method was compared with the conventional method in which these parameters are fixed to default values. The simulation conditions are shown in Table II-9.4-1.

Table II-9.4-1 Simulation Conditions

The number of gNodeBs, cells	1, 1
Frequency, band	4.8 [GHz], n79
Bandwidth	100 [MHz]
Subcarrier spacing (SCS)	30 [kHz]
Duplex	TDD
Downlink to Uplink ratio	1:1
Transmission power	23 [dBm]
Floor area	100 [m] x 100 [m]
Floor layout	layout assuming a factory
The number of simultaneously running robots	up to 18
Robot running speed	up to 3 m/s
Traffic per robot	downlink: up to 150Kbps uplink: up to 1 Mbps

Fig. II-9.4-1 shows the simulation results regarding the relative frequency of unmet QoS requirements per packet in an environment where both QoS requirements and radio quality fluctuate based on field conditions, such as driving speed and surrounding circumstances. Our method significantly reduced the number of packets failing to meet QoS requirements, achieving less than 1/50 compared with the conventional method. In other words, the number of system outages due to communication issues in mobile robot operations was significantly decreased, effectively improving the mean time between failures by a factor of 50.

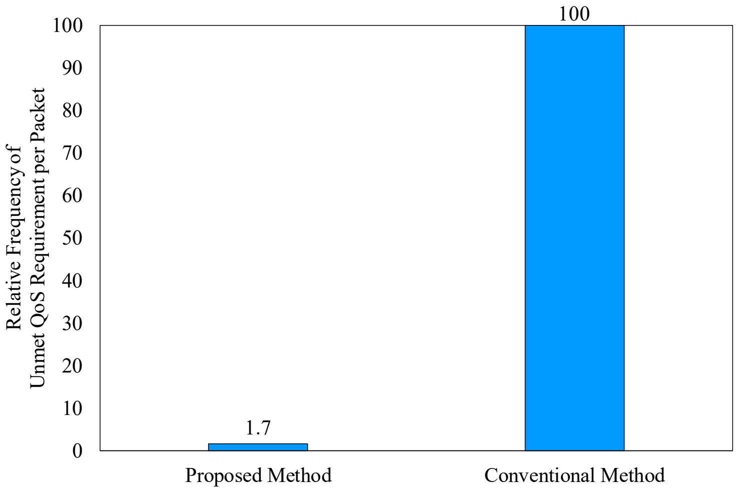


Fig. II-9.4-1 Simulation Results

II-9.5. Future Vision

Conventionally, IoT devices are equipped with intelligent functions specific to their vendor or model, and the IoT controller software is also tied to a particular vendor and model. When AI (for RAN) and 5G/6G realize an adaptive and 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 IoT controller installed in the cloud or edge server to control IoT devices of multiple vendors coordinately and for various models to optimize the entire system. Furthermore, achieving simplification, lightweight implementation, and generalization of IoT devices will likely drive the spread of IoT solutions and, as a result, accelerate the developments in IoT applications and AI (for IoT).

II-9.6. Conclusion

Mobile network specifications will become more sophisticated in the 5G/6G era. However, intelligent network optimization during operation will be essential for adapting to the evolving conditions of applications, networks, and sites. An application-aware RAN optimization method based on Open RAN architecture was introduced to ensure strict QoS requirements across various vertical domains while accommodating diverse application needs and fluctuations in wireless quality. The simulation results of applying the proposed method to a system that remotely controls multiple autonomous robots operating in factories/warehouses confirmed that the number of robot stoppages could be reduced by 98% or more compared to the scenario where the method was not utilized.

Advancements in AI (for RAN) and 5G/6G technologies aim to deliver adaptive and reliable wireless communication that meets the QoS requirements of various applications. These improvements will facilitate sophisticated and high-load data processing on cloud or edge servers. Additionally, they will enhance the management and optimization of IoT devices from diverse vendors, simplifying these devices to accelerate the development of AI (for IoT) and IoT applications. This, in turn, will promote the broader adoption of IoT solutions.

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] O-RAN "O-RAN Empowering Vertical Industry: Scenarios, Solutions and Best Practice," White Paper, Dec. 2023.

[3] 5G ACIA, "Key 5G Use Cases and Requirements," White Paper, May 2020.

[4] 5GAA, "C-V2X Use Cases Volume II: Examples and Service Level Requirements," Oct. 2020.

[5] O-RAN ALLIANCE, https://www.o-ran.org/

[6] O-RAN Working Group 2, "AI/ML workflow description and requirements," Technical Report v1.1, 2020.

[7] 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

[8] 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

II-10. AIOps for Autonomous Network

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.**

II-10.1. 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. II-10.1-1, 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.

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. II-10.1-1, 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.

II-10.2. AIOps and Autonomous Network

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.

Furthermore, with the remarkable development of LLM technology in recent years, it is expected that R&D and standardization of Agentic AI, in which AI Agents with LLM functions autonomously provide instructions and reports on each task while interacting with human operators, will accelerate over the coming years.

II-10.3. Network Failure Management by AIOps

Our proposed integrated framework [4] for network failure management with AIOps is shown in Fig. II-10.3-1, 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.

II-10.4. 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.

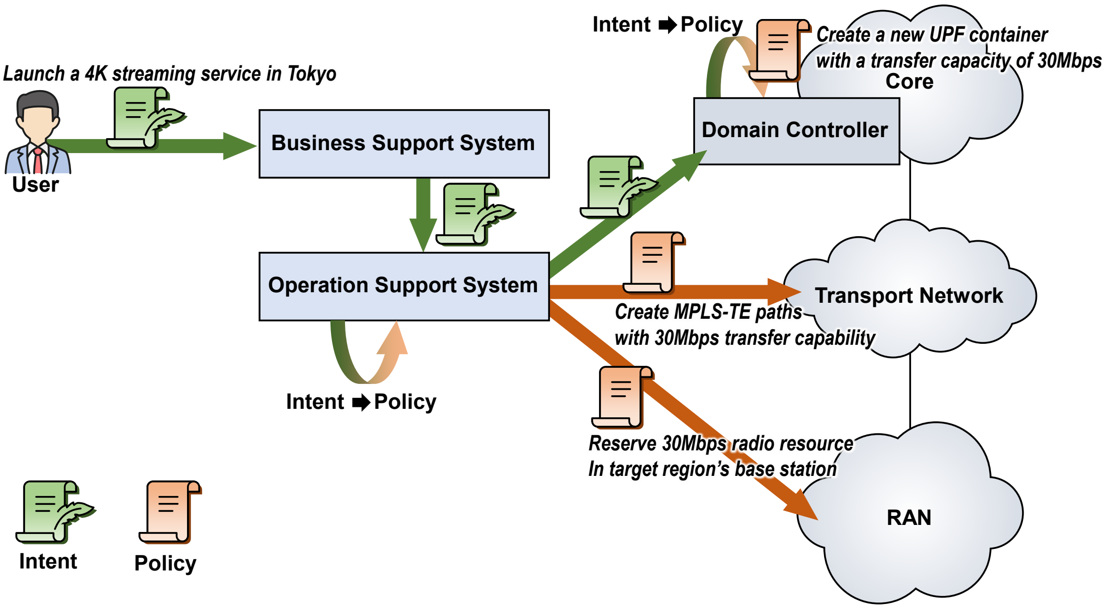

Intent ➡ Policy
Create a new UPF container
with a transfer capacity of 30Mbps
Core
Launch a 4K streaming service in Tokyo
User
Business Support System
Domain Controller
Operation Support System
Transport Network
Create MPLS-TE paths
with 30Mbps transfer capability
Intent ➡ Policy
Reserve 30Mbps radio resource
In target region's base station
RAN
Intent
Policy

Fig. II-10.1-1 General concept of Autonomous Network

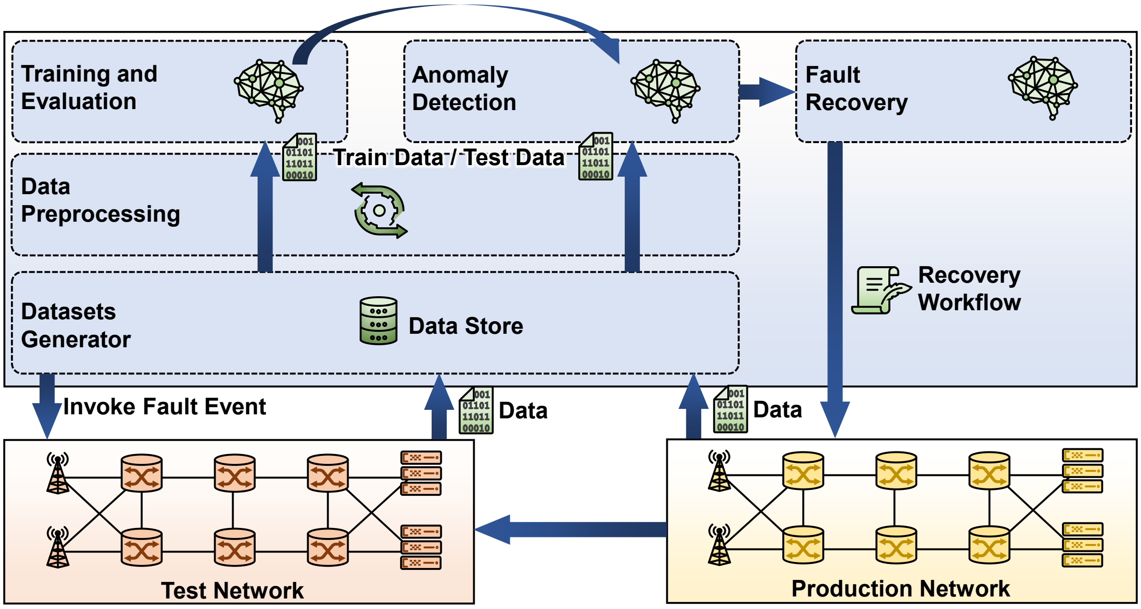


Fig. II-10.3-1 AIOps framework for Network Failure Management

REFERENCE

[1] A. Boasman-Patel, et al., “Autonomous Networks: Empowering Digital Transformation forf 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.

[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.

II-11. 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 propose a logic-oriented generative AI for intent translation, which is a logical search engine enhanced by AI/ML technology. This paper presents the position of our proposal with respect to related techniques, and then briefly outlines its method.**

II-11.1. Introduction

Modern networks continue to grow in complexity. Their rapid and stable provision is becoming increasingly difficult, and a high degree of operational automation is required [1][2]. Intent-based networking is one of the promising foundational approaches 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. This translation corresponds to the design of the network and requires complex logical thinking. 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. However, the former has a problem with the flexibility of possible answers. 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.

II-11.2. Automation of Intent Translation and Its Challenges

Intent-based networking is a new technology that provides an abstraction layer for network control [4]. It allows users to control the network by directing the desired state of network services instead of telling them how to configure network services. There are various issues 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, the means to derive concrete network configuration from abstract Intent, i.e., translation, is a key issue in realizing Intent-based networking. Figure II-11.2-1 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.

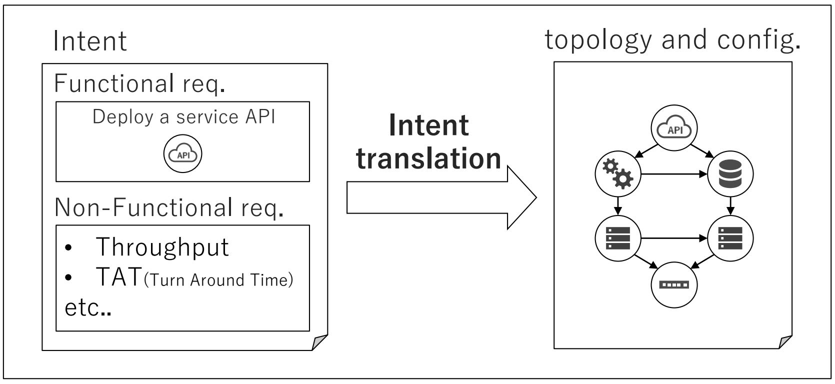


Fig. II-11.2-1 Concept of automated network intent translation.

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) 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]. Thus, a deductive approach has faithfulness but lacks flexibility, while an inductive approach has extremely high flexibility but has problems with faithfulness.

In response to this situation, in the area of LLM, a method to increase logical accuracy by dividing thinking into detailed steps has been proposed in recent years, and a number of services are already available. This is a method that adds a deductive element to inductive methods and improves logical accuracy while maintaining flexibility. On the other hand, this paper proposes a method to improve flexibility while maintaining logical validity by adding inductive elements to the deductive method described above.

II-11.3. Intent Translation with Logic-oriented Generative AI

The left side of Figure II-11.3-1 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 Figure II-11.3-1. 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.

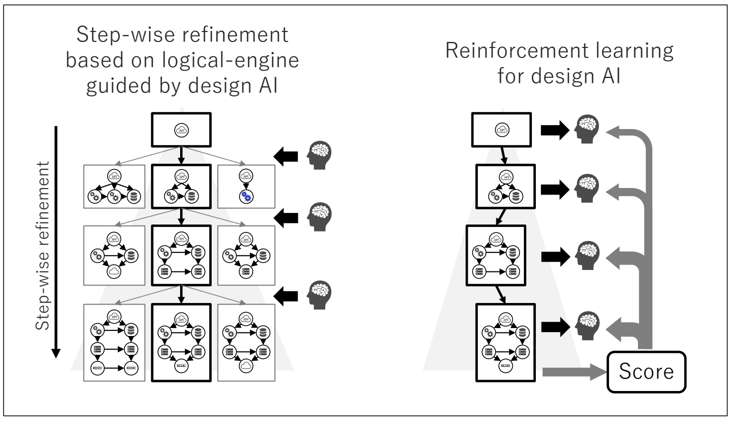


Fig. II-11.3-1 concept of automated intent translation and its learning.

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.

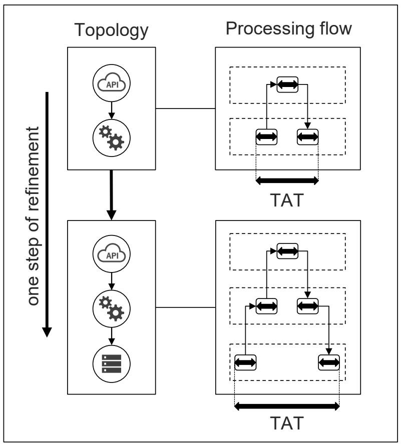


Fig. II-11.3-2 Example of one step refinement of network intent and evaluation model.

The validity of the design results of this technique is guaranteed by the constraints. Figure II-11.3-2 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 Figure II-11.3-2 shows, TAT cannot be accurately determined until the last step. The incomplete processing flow shown in the upper right corner of Figure II-11.3-2 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.

II-11.4. 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-of-thought reasoning”, 2023. arXiv:2301.13379.

[7] https://doi.org/10.48550/arXiv.2301.13379

II-12. In-Network Learning for Distributed RAN AI, ~Distributed LLMs via Latent Structure Distillation~

Takashi KOSHIMIZU
Huawei Technologies Japan
Chenghui PENG, Shaoyun WU
Huawei Technologies

***Abstract*— This paper proposes a distributed learning algorithm, named In-Network Learning (INL) for inference over wireless radio access networks (RANs) without transmitting raw data. The applied algorithm is suitable for both *multimodal* and *heterogeneous* data where the fusion of features extracted in a distributed manner. It also offers substantial gains over state-of-the-art (SOTA) solutions such as Horizontal and Vertical Federated Learning (FL) and Horizontal and Vertical Split Learning (SL) in terms of both accuracy and bandwidth requirements. This eventually discuss how the algorithm can be extended to support the deployment of LLMs and knowledge distillation in wireless networks.**

II-12.1. Distributed inference over Wireless RANs

RANs have important intrinsic features that may pave the way for cross fertilization between machine learning (ML) and communication. This is in contrast to simply replacing one or more communication modules by applying ML algorithms as black boxes. While relevant data is generally available at one point in areas such as computer vision and neuroscience, it is typically highly distributed across several sites in wireless networks. Such examples also include channel state information (CSI) and/or the so-called radio-signal strength indicator (RSSI) of a user's signal, which can be used for things like localization, precoding, or beam alignment [1].

Conventional common approach for implementing ML solutions involves collecting all relevant data at one site (e.g., a cloud server or macro BS) and then training a suitable ML model using all available data with ample processing power. However, this approach may not be suitable in many cases due to large data volume size and scarcity of resources, including power and bandwidth consumption. Additionally, some applications (e.g., automatic vehicle driving) asked stringent latency requirements that are incompatible with data sharing. In addition, it might be desirable not to share the raw data in order for user privacy. Furthermore, edge devices such as small BSs, on-board sensors, and UEs typically have limited memory and computational power. Also, the wireless environments are typically prone to change rapidly, e.g., fluctuate connectivity and occasional joining/leaving devices. Data dynamics is another criticality, where the data become more multimodal and heterogeneous across devices and users. Table-II-12.1-1 summarizes the main features of inference over wireless RANs.

Table II-12.1-1. Summary of the main features of inference over wireless RAN

Data	Inference	Network connectivity/topology	Privacy	Compute resources
• Distributed during training • Distributed during inference • Multimodal • Heterogeneous	• Distributed • Communication is the bottleneck • Fusion is required • Extremely short latency (≈ 0.1 ms)	• Prone to rapid changes: - Users joining or leaving the network - Link failure - Channel quality drop-off	• Critical - Raw data leaks information about users	• Small • Distributed across sites

II-12.1.1. AI at the Wireless Edge

The challenges above mentioned have faced a new paradigm called "edge learning" and/or distributed learning, where intelligence moves from the center to its network edges. In such scenario, the system design plays a central role because both data and computational resources are highly distributed. The goal of distributed inference over RAN is to make decisions on one or more tasks, at one or more sites, by exploiting the available distributed data. In this framework, multiple devices (e.g., BSs and UEs) are each equip with a neural network (NN). Some of the devices possess data they have acquired through communication or sensing, whereas some only contribute to the collective intelligence through computational power, as in Fig. II-12.1.1-1.

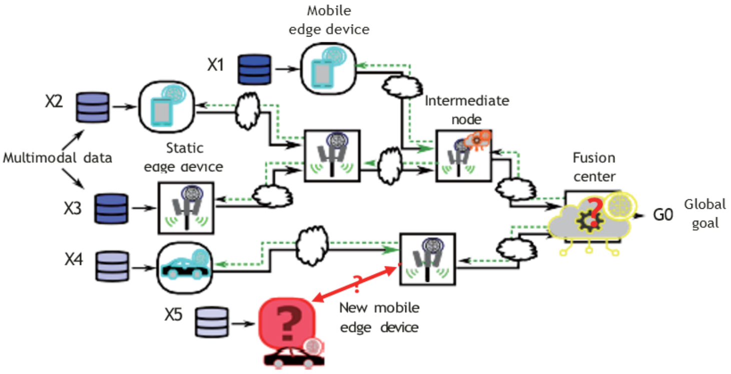


Fig. II-12.1.1-1 Distributed inference over RAN

II-12.1.2. Brief Review of SOTA Algorithms

AI solutions for RANs can be classified according to whether only the training phase is distributed (such as Horizontal Federated Learning and Horizontal Split Learning) or both the training and inference (or test) phases are distributed (such as Vertical Federated Learning).

- **Horizontal Federated Learning (HFL):** HFL would be the most popular distributed learning scheme [2]. It is considered most suitable for settings in which the training phase is performed in a distributed manner while the inference phase is performed centrally. During the training, each client equips a distinct copy of a same NN model where the client trains on its local dataset. The learned weight parameters are then sent to and aggregated by (e.g., their average is computed) a cloud server or parameter server (PS). This process is repeated, each time using the obtained aggregated model for reinitialization, until a convergence found. The advantage of this approach ensures the model is progressively adjusted to account for all variations of the data, not only those of the local dataset.

- **Vertical Federated Learning (VFL)** [3]: VFL is a variation of FL, the data is partitioned vertically and both the training and inference phases are distributed. Fig. II-12.1.2-1 illustrates the data structure in HFL and VFL respectively. In this case, client device holds whole data that is relevant for a possibly distinct feature. A prominent application example can be seen where the data is heterogeneous across clients and/or multimodal. In VFL, different clients may apply distinct NN models that are tailored for their own data modalities. These models are trained jointly to extract features that are collectively enough to make a reliable decision at the fusion center, as in Fig. II-12.1.2-2a. For recent advances on VFL and its applications in wireless settings can be also find in [4, 5] with the references.

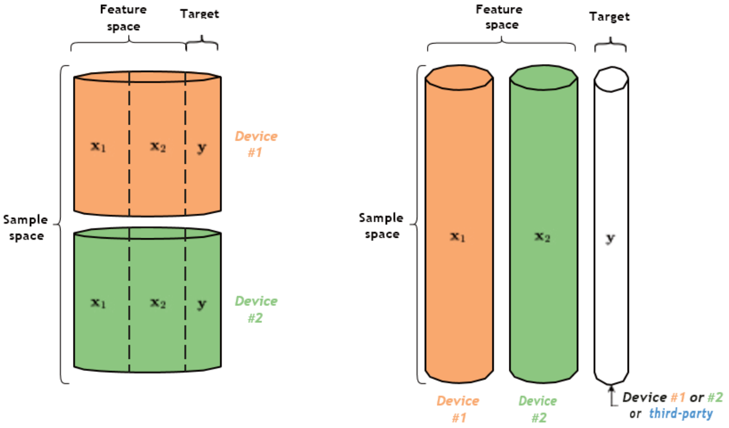


Fig. II-12.1.2-1 HFL (left) and VFL (right) structure

- **Split Learning (SL)** [6]: Similar to FL, it has two variations: Horizontal SL (HSL) and Vertical SL (VSL). VSL was introduced earlier than VFL, now it can be viewed as a special case of a VFL. For HSL, a two-part NN model is split into an encoder part and a decoder part. Each edge device possesses a copy of the encoder part and both NN parts can be learned sequentially. The decoder does not have its own data, whereas in every training round, the NN encoder part is fed with the data of one device and its parameters are initialized using those learned from the previous round. Then, during the inference phase, the learned two-part model is applied to centralized data.

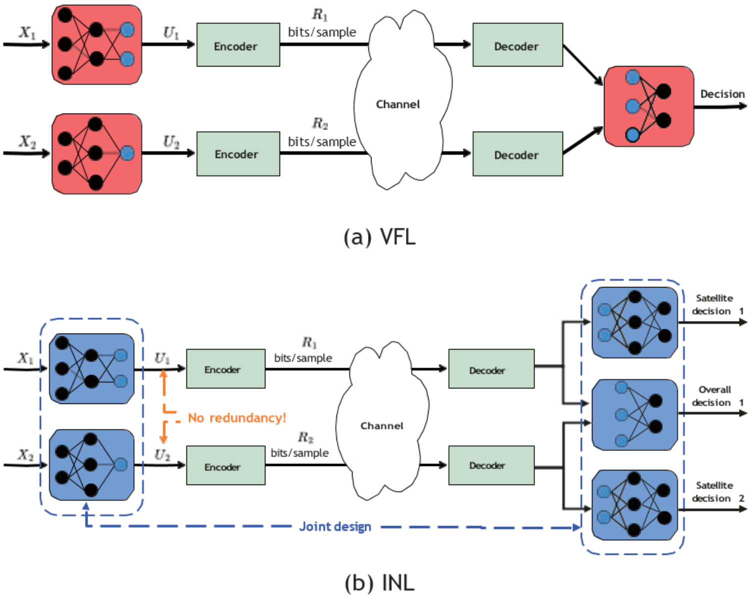


Fig.II-12.1.2-2 Feature redundancy removal by INL

II-12.2. In-Network Learning

The roots of In-Network Learning (INL) can be seen in [7, 8], with further development have been also taking place in [9–11]. INL is the most expected ML scheme for distributed inference in heterogeneous and multimodal data. This scheme assuming every device equips an NN. During the inference process, each device independently extracts suitable features from its input data for a given inference task. These features are then transmitted over the network and converged at a given fusion center in order to obtain a reliable decision. These devices that hold useful data (these devices play the role of encoders) perform individual feature extraction independently from each other. Through the training, algorithm ensures that the encoders only extract complementary features, for instance, redundant inter-device features are removed, which enabling substantial bandwidth savings. The key technical characteristics in this algorithm are listed as follows:

- **Network Feature Fusion:** INL fuses features that are extracted in a distributed manner at a fusion center so, collectively they enable a desired decision to be made at the fusion center after being transmitted over the network.
- **Feature Redundancy Removal:** A distinguishing factor of INL is that, during inference, the encoders only extract non-redundant features and they are trained

during training phase. Specifically, during inference process, each encoder only extracts features that are useful for a given inference task from its input-data while also considering the other features extracted by other encoders.

- **Feature Extraction Depends on Network Channel Quality**: Encoder feature extraction also considers the quality of the channel to the fusion center. Hence, the features are extracted only to the extent that is possible to transmit them reliably to the decision maker.
- **Satellite Decoders**: The fusion center is equipped with a main decoder and satellite decoders, which are trained to make soft decisions based on the individual features transmitted by the encoders, the system is depicted in Fig. II-12.1.2-2b.

II-12.3. Preference Gains

This section compares the algorithm performance on INL versus HFL and HSL in terms of achieved accuracy and the bandwidth requirements.

Experiment 1: We prepare five-sets of noisy versions of images obtained from the CIFAR-10 dataset [12]. The images are first normalized, and then corrupted by additive Gaussian noise with standard deviation (σ) is set respectively to 0.4, 1, 2, 3, 4. For INL, each of the five input NNs are trained on a different noisy version of the same image. Each NN uses a variation of the VGG network of [13], with the categorical cross-entropy as the loss function. The architecture is shown in Fig. II-12.3-1. In the experiments, all five noisy versions of every CIFAR-10 images are processed simultaneously, each by a different NN at a distinct node.

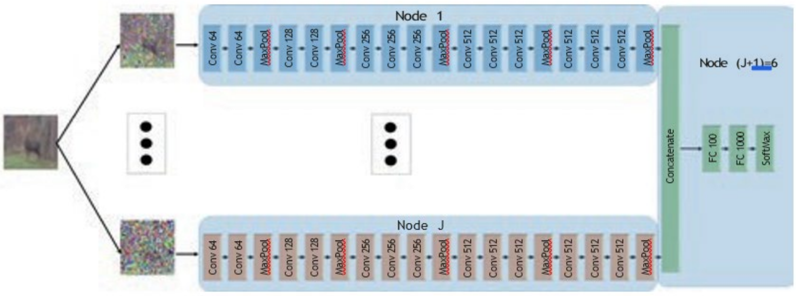


Fig. II-12.3-1, NW architecture. Convolutional layer & Fully connected layer

Subsequently, the outputs are concatenated and then passed through a series of fully connected (FC) layers at node (J + 1). For HFL, each of the five client nodes is equipped with the entire network of Fig. II-12.3-1. The dataset is split into five sets of equal sizes, with the split being performed such that all five noisy versions of a given CIFAR-10 image are presented to the same client NN (note: however, that distinct clients observe different images).

For HSL, each input node is equipped with an NN formed by all five branches with convolution networks (i.e., the entire network shown in Fig. II-12-3-1, except the part at Node (J + 1)). Furthermore, node (J + 1) is equipped with fully connected layers at Node (J + 1). Here, the processing during training is such that each input NN vertically concatenates the outputs of all convolution layers and then passes that to node (J + 1), which then propagates back the error vector. After one epoch at one NN, the learned weights are passed to the next client, which performs the same operations on its part of the dataset.

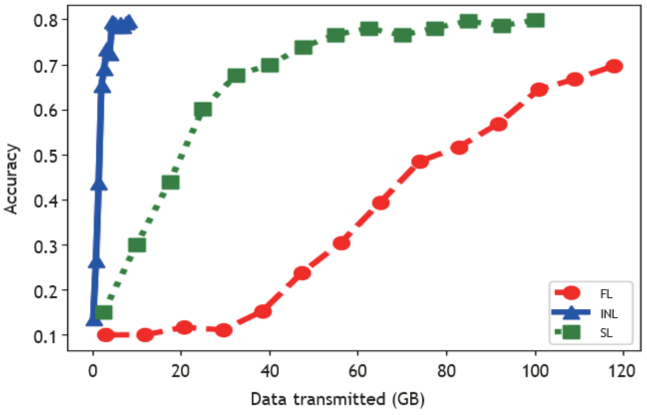


Fig. II-12.3-2 Accuracy vs. bandwidth cost for Experiment-1

Fig. II-12.3-2 shows the amount of data needed to be exchanged among the nodes (i.e., bandwidth resources) in order to get a prescribed value of classification accuracy. It can be observed that our INL requires significantly less data transmission than HFL and HSL for the same desired accuracy level.

Experiment 2: In the previous experiment, the entire training dataset was partitioned differently for INL and HFL in order to account for their unique characteristics. In the second experiment, they are all trained on the same data. Specifically, each client NN sees all CIFAR-10 images during training, and its local dataset differs from those seen by other NNs only by the amount of added Gaussian noise (with σ is set respectively to 0.4, 1, 2, 3, 4). Also, to ensure a fair comparison of the three schemes, INL, HFL, and HSL, we set the nodes to utilize the same NNs fairly for each of them in Fig. II-12.3-3.

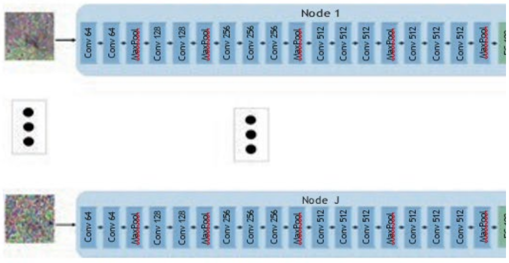


Fig.II-12.3-3 Used NN architecture for Experiment-2

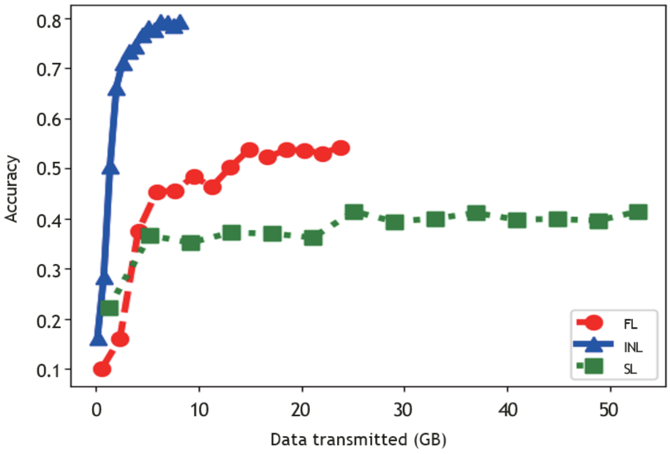


Fig.II-12.3-4 Accuracy vs. bandwidth cost for Experiment-2

Fig. II-12.3-4 shows the performance of the three schemes during the inference phase in this experiment. For HFL, the inference is performed on the image whose quality is the average among the five noisy input images used for INL. Again, it can be observed that the benefits in INL over HFL and HSL in terms of both achieved accuracy and bandwidth requirements.

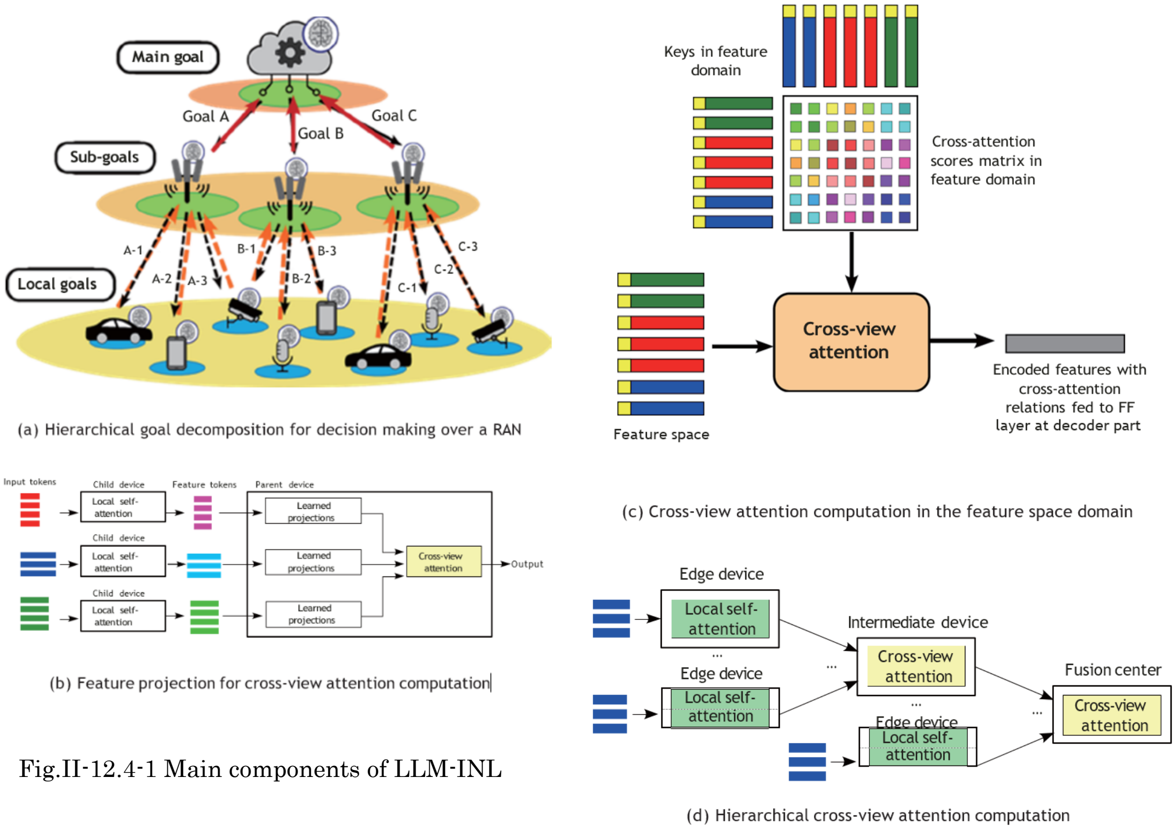
II-12.4. LLM

In addition to having remarkable capabilities, LLMs are significantly contribute overall AI development and even re-shaping our future. However, their multimodality, in part, causes some critical challenges in the cloud-based deployment: (i) response time, (ii) communication bandwidth cost, and (iii) infringement of data privacy. Therefore, an urgent need identified to leverage Mobile Edge Computing (MEC) in order to finetune and deploy LLMs on or in closer proximity to data sources, while also preserving data ownership for end users. In accordance with the vision of "NET4AI" (network for AI) in 6G era [15], we envisioned a 6G-MEC architecture that can support LLM deployment at the network edge. Our proposed architecture includes the following critical modules.

- **Goal Decomposition**: The global inference task is performed collaboratively between different layers in the mobile network system. The fusion center decomposes the global goal into smaller sub-goals and assigns them to the next-layer BSs based on their respective strengths. The BSs then further decompose the sub-goals into smaller ones. This process continues until it reaches the edge devices, as in Fig. II-12.4-1a.
- **Cross-View Attention**: The self-attention of transformers can only be computed for locally available sensory data. If multiple sensors acquire multi-view data that is relevant for a given inference task, it is necessary to compute how a token from a given piece of data collected at one sensor attends to another token from another piece of data collected or measured at another sensor. We call this as cross-view

attention, which is computed at a fusion center in the feature space after feature projection on a hyperplane, as in Fig. II-12.4-1b, II-12.4-1c, and II-12.4-1d.

- **Latent Structure-based Knowledge Distillation**: It is expected that 6G will evolve into with mobile network supporting in-network and distributed AI at the edge [15]. However, considering the excessive memory and compute requirements of LLMs, is it feasible to run such large models at the 6G edge? Also, would the network bandwidth support various agents/devices equipped with LLMs exchanging the entirety of their models for model aggregation and collaboration? A step in this direction has been studied in [16] recently, where devices use INL to only exchange the structure of their extracted features, not the features themselves. This structure is then utilized onsite at the device to fine-tune the locally extracted features.



(a) Hierarchical goal decomposition for decision making over a RAN

(c) Cross-view attention computation in the feature space domain

(b) Feature projection for cross-view attention computation

Fig.II-12.4-1 Main components of LLM-INL

(d) Hierarchical cross-view attention computation

II-12.5. Conclusion

This paper explained our proposal and analysis on INF for the inference application for AI native 6G cellular network. The performance evaluations are also examined specifically on INL comparing that in HFL and HLS. It also explained LLM application in INL. The full set of original paper on this contribution can be seen in [17].

REFERENCE

[1] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI- based fingerprinting for indoor localization: A deep learning approach," IEEE Transactions on Vehicular Technology, vol. 66, no. 1, pp. 763–776, 2017.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial Intelligence and Statistics. PMLR, 2017, pp. 1273–1282.

[3] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen and T. Ranbaduge, "Vertical federated learning: Challenges, methodologies and experiments," arXiv preprint arXiv: 2202.04309, 2022.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp.1–210, 2021.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[6] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," Journal of Network and Computer Applications, vol. 116, pp. 1–8, 2018.

[7] I. E. Aguerri and A. Zaidi, "Distributed Variational Representation Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 120–138, 2021.

[8] I. Estella Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in International Zurich Seminar on Information and Communication (IZS 2018). Proceedings. ETH Zurich, 2018, pp. 35–39.

[9] M. Moldoveanu and A. Zaidi, "In-network Learning for Distributed Training and Inference in Networks," in 2021 IEEE Globecom Workshops (GC Wkshps). IEEE, 2021, pp. 1–6.

[10] --, "On in-network learning. A comparative study with federated and split learning," in 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, 2021, pp. 221–225.

[11] --, "In-network learning: Distributed training and inference in networks," Entropy, vol. 25, no. 6, p. 920, 2023.

[12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[15] L. Huawei et al., "6G: The next horizon from connected people and things to connected intelligence," Huawei, White Paper, 2022.

[16] M. Sefidgaran, A. Zaidi, and P. Krasnowski, "Minimum description length and generalization guarantees for representation learning," Advances in Neural Information Processing Systems, vol. 36, 2023.

[17] A. Zaidi, et al., “In-Network Learning for Distributed RAN AI, ~Distributed LLMs via Latent Structure Distillation~”, Communications of HUAWEI RESEARCH, November 2024.

II-13. Throughput Prediction Technology for 28 GHz Channels using Physical Space Information

Takahiro Yamazaki, Hisashi Nagata, Riichi Kudo,
Kahoko Takahashi, Takayuki Yamada, Takafumi Fujita
NTT Network Innovation Laboratories, NTT Corporation

Abstract— Advances in wireless communications, such as the 5th-generation mobile communication system (5G), have enabled a wide variety of devices to be connected to wireless networks. In 6G, all physical entities will be connected to wireless networks and their physical space information, such as position and velocity, will be available for new mobile services. NTT's Innovative Optical and Wireless Network (IOWN) will accelerate to obtain the physical-space information from various sensors. Therefore, mobile traffic is growing rapidly toward 6G. The use of the millimeter-wave (mmWave) bands is promising to increase the capacity of mobile networks. However, the mmWave link quality (LQ) is strongly affected by surrounding objects. To stably use mmWave bands, an effective solution is to predict future LQ and adaptively control wireless communication. This article introduces 5G throughput-prediction technology that is based on deep neural networks using physical-space information and an automated 5G measurement environment using humanoid robots for deep-learning evaluations.

II-13.1. Background and Overview

Advanced wireless communication systems enable a wide variety of devices connect to wireless networks. The 5th-generation mobile communication system (5G) contributes to creating a wide range of innovative applications, such as virtual and augmented reality (VR/AR), as well as services in diverse industries that require high speed, low latency and high reliability [1]. In 6G, all elements including people, things, and systems, will be connected to wireless networks, and an advanced cyber-physical fusion system (CPS) is expected to feedback optimal results to the real world through artificial intelligence (AI) [2]. A CPS is a system concept in which AI creates a replica of the real world in cyberspace (digital twin) and emulates it beyond the constraints of the real world. This concept will provide various values and solutions to social problems. NTT's Innovative Optical and Wireless Network (IOWN) [3] accelerates the CPS concept by collecting physical space information from all devices and generating big data; thus, mobile traffic is growing rapidly toward 6G. The compound annual growth rate of mobile data usage worldwide is reported to 60 % [4].

In order to accommodate the explosion in mobile traffic, the use of higher frequencies such as millimeter-wave (mmWave) is key for future wireless communication systems [5]. However, the mmWave bands are characterized by strong direct wave radio propagation, and the mmWave link quality (LQ) is strongly affected by surrounding objects. To ensure stable use of the mmWave bands, we introduce wireless LQ prediction technology that focuses on the relationship between physical-space information and wireless link information. NTT Network Innovation Laboratories is researching using physical-space information to promote the evolution to wireless communication systems toward IOWN/6G [6].

II-13.2. System Model of Wireless LQ Prediction

Recent study of wireless LQ prediction for mmWave bands showed that physical space information such as user equipment (UE) position, camera images and point cloud data, are strongly correlated with wireless LQ of mmWave. For example, the received signal-strength-indicator (RSSI) prediction for 60 GHz using depth images from RGB-D cameras in an indoor environment where two pedestrians move between an access point and fixed UE has been investigated [7]. However, a more complicated and practical scene where both the UE and surrounding objects move has not been considered. Therefore, we developed the wireless LQ prediction system for the complicated scene [8].

Fig. II-13.2-1 illustrates a wireless LQ prediction system that predicts future wireless LQ using physical-space information. The system assumes an environment where pedestrians walk around in a wireless cover area of a base station. There are two types of pedestrians, one is a UE holder, and the other is a pedestrian for blocking. The UE holder walks while accessing applications such as VR/AR through a base station. The pedestrian for blocking has no UE and just walks around the UE holder. The system uses cameras/sensors to gather physical space information, which are the position, direction, and velocity of all the objects such as the pedestrians. The system also gathers wireless LQ information such as data thruput from the UE. The LQ prediction model is trained with the physical space information and the wireless LQ information by using machine learning algorithms.

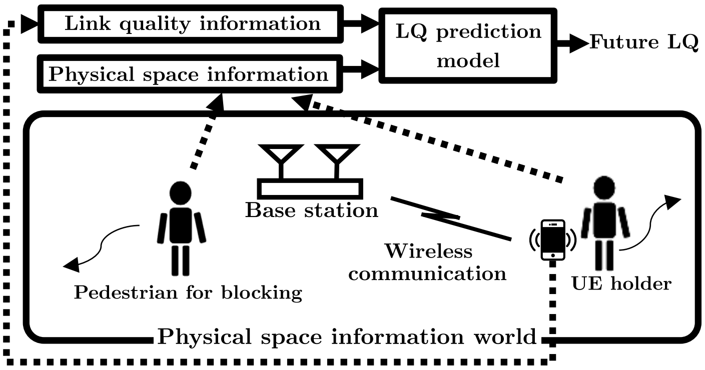


Fig. II-13.2-1. System model of wireless LQ prediction.

II-13.3. Experimental Evaluation of the Wireless LQ Prediction System

To evaluate the wireless LQ prediction system that we described in the previous section, we gathered training data in an indoor environment with 28 GHz 5GNR and experimentally evaluated a prediction accuracy of the LQ prediction model [8]. We implemented a deep neural network (DNN) for the LQ prediction model and used a communication throughput as the LQ.

For the experimental evaluation, we considered a pedestrian scenario in which two people are walking around in an indoor room; one is the UE holder who has a UE which communicates via a 5G 28 GHz channel, and the other is the pedestrian for blocking. For this scenario, we built autonomous mobility humanoid robots to gather an enough amount of training data for the LQ prediction model. The humanoid Robots-A and -B were used as substitutions for the UE holder and the pedestrian for blocking, respectively. Fig. II-13.3-1 shows an indoor experiment map and the different routes of the two robots. The running routes of Robot-A with the UE and Robot-B were the red and green lines, respectively, in this figure. Each robot flipped at the ends of the line and continued going back and forth between the ends of the line. The maximum robot speed was 1.0 m/s. Robot-B ran between Robot-A and the base station, resulting in a decrease in throughput due to blocking. Each robot consisted of a humanoid mannequin mounted on a mobility robot. Robot-A, which held the UE in a backpack is shown in Fig. II-13.3-2. Robot-B for blocking is shown in Fig. II-13.3-3. Robots-A and -B were 1.67 and 1.70 m tall, respectively. The antenna height of the base station is 2.65 m. These robots were controlled by a robot operating system (ROS) [9]. The robots' position, velocity, and direction were obtained from the ROS as physical space information. The robots had LiDAR (light detection and ranging) censors which can collect point clouds of laser signal reflection points. The point clouds were used to calculate the location and direction of the robots.

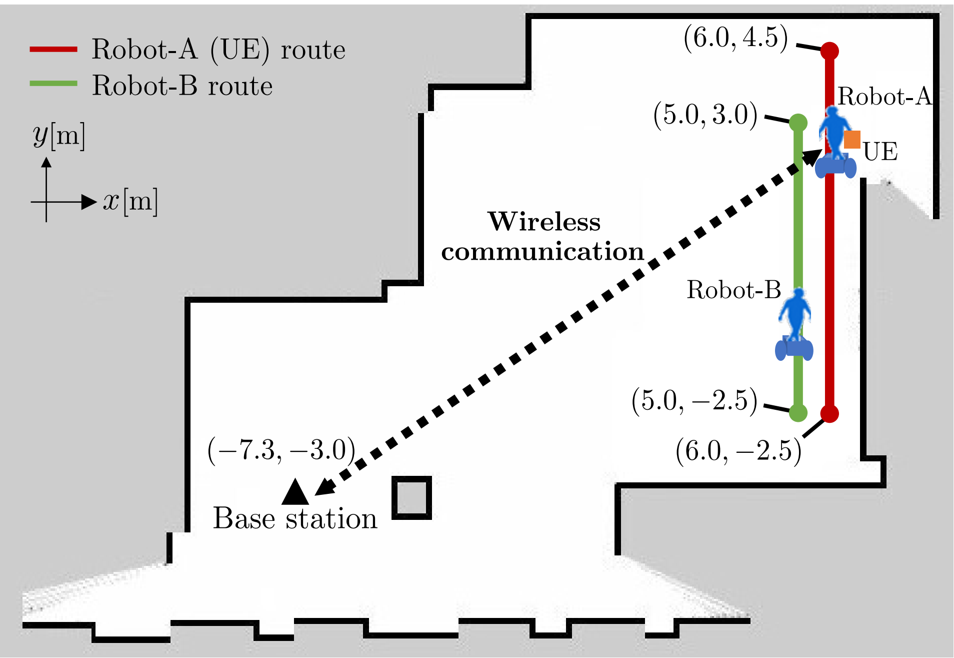


Fig. II-13.3-1. The locations of UE and 5G base station are shown on a map of the indoor experimental environment. The running courses of Robot-A and -B are red and green lines, respectively. Each robot flips at the end of the line.

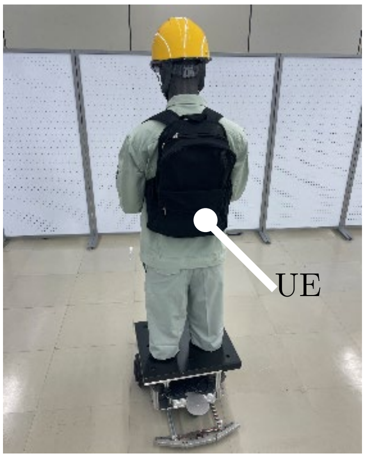


Fig. II-13.3-2. Robot-A with UE communicates through the 5G base station.

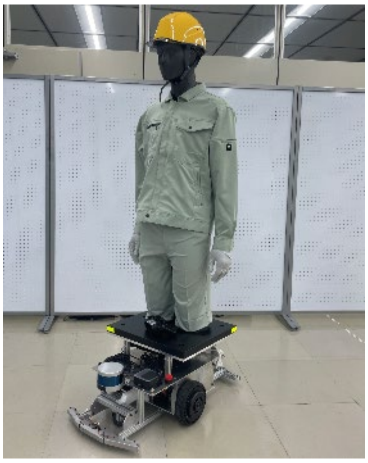


Fig. II-13.3-3. Robot-B for blocking runs between Robot-A and the 5G base station.

We used the Iperf [10] software tool to measure the throughput of the UE. In this experiment, we focused on an uplink communication, so we made the UE transmit packets via the Iperf to the server which was set on the multi access edge computing (MEC).

The throughput of the UE and the states of the Robots-A and -B were measured every 100 ms. The resulting dataset contained 1,493,750 samples corresponding to about 41 hours spread over 11 days. These values were normalized to yield a distribution from 0 to 1 or -1 to 1. We used 20 % sampling data as test data. The remaining 80 % was used as training data (90 %), and as a validation data (10 %). Since we focus on LQ prediction for detecting the performance drops due to shielding, the upper limit of the measured throughput was set to 200 Mbps. Therefore, our prediction scheme predicts the future throughput with ceiling of 200 Mbps. For the LQ prediction model, we used a DNN which has three hidden layers: one long short-term memory layer and two fully connected layers spaced with 10% dropout. The activation function for the hidden layers is the rectified linear unit. The DNN is trained to minimize the loss function of the mean squared error. The optimization algorithm is Adam with a learning rate of 0.0005. The output value of the DNN is one-second-ahead throughput. To evaluate the differences in prediction accuracy of the DNN due to the input features of training data, we prepared four input features: one is the past one-second throughput (Φ_{T}), one is the past one-second states of the Robot-A (Φ_{A}), one is the past one-second states of the Robots-A and -B (Φ_{AB}) and one is the past one-second throughput and states of the Robots-A and -B (Φ_{ABT}).

Fig. II-13.3-4 shows time sequential plots of one-second-ahead throughput prediction values and measured throughput values. There were two main factors affecting throughput degradation in our scenario. The first was line-of-sight (LOS) blockage by Robot-B moving between Robot-A and the base station at around 26 and 58 seconds, as shown in Fig. II-13.3-4. The observed throughput dropped to about 100 Mbps due to the blocking effect of the robot body in our environment. This occurred at various locations along Robot-A's route, and blocking time changed due to the speed and relative directions of Robots-A and B. The second factor was self-blocking by Robot-A, which made a 180

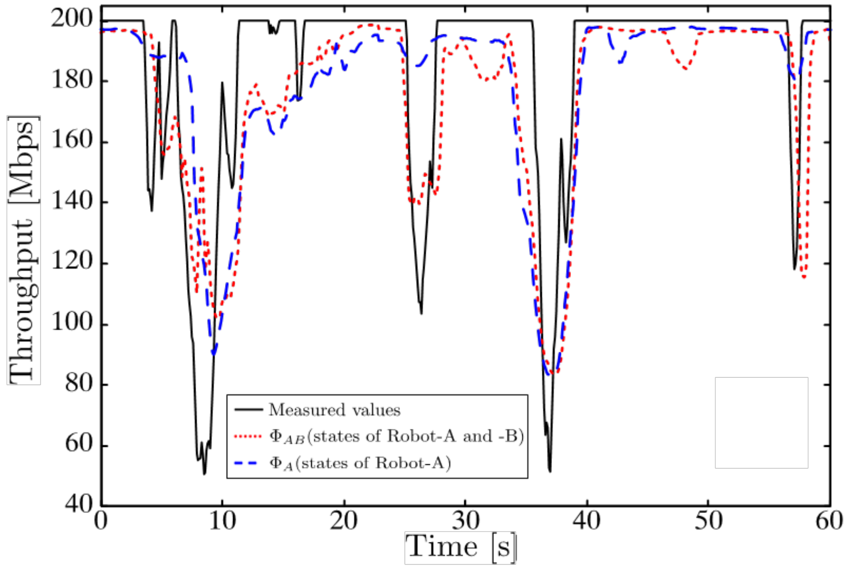


Fig. II-13.3-4. Time sequential plots of measured and predicted

degree turn at each endpoint goal at around 9 and 37 seconds, as shown in Fig. II-13.3-4. The self-blocking by Robot-A had greater impact on throughput than the LOS blockage by Robot-B as the throughput rapidly dropped below 50 Mbps. This is because Robot-A as the obstacle (self-blocking) was closer to the UE than Robot-B, and the 180-degree turn took longer to complete than the blockage by Robot-B.

Fig. II-13.3-5 shows the cumulative distribution function (CDF) of the absolute error between the predicted throughput values and measured throughput values. The effectiveness of physical-space information became more prominent as the CDF value fell. The 50th-percentile absolute error value improved by 57.5% using Φ_{ABT}, compared with using Φ_{T}, which takes past throughput as the input feature. This result indicates the correlation between the physical-space information and throughput. Additionally, the 70th-percentile absolute error value was less than 20 Mbps for all input features, indicating that the absolute errors were concentrated within 20 Mbps and correlations were observed between all input features and throughput. Similarly, at the 50th-percentile absolute error value, compared with Φ_{A} and Φ_{AB}, an improvement of 35% was attained by adding the state of Robot-B. This confirms the effectiveness of the states of surrounding objects, such as Robot-B, in throughput prediction. Fig. II-13.3-5 also shows that large prediction errors exceeding 50 Mbps occurred. This is because the current input features of past throughput and physical-space information cannot explain the 5G network-driven throughput changes such as link interruption and reconnection. For future work, we plan to consider such throughput changes by adding the 5G network information.

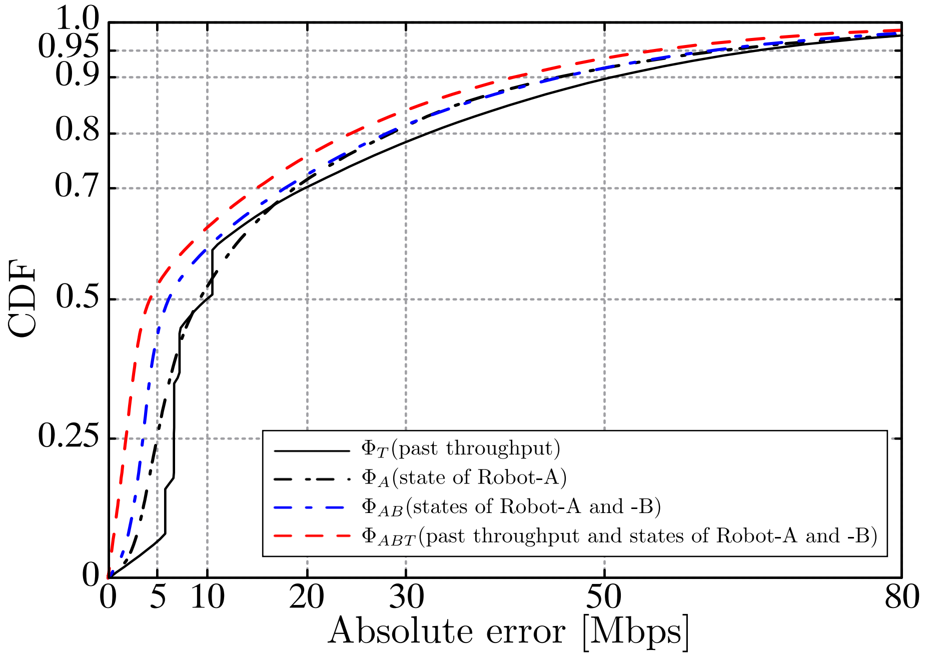


Fig. II-13.3-5. CDF of the absolute error between the predicted throughput for each input feature and measured throughput.

II-13.4. Conclusion

This article presented a throughput-prediction technology for 5G services over a 28-GHz channel that uses physical-space information for a two-pedestrian scenario in which both a UE holder and a pedestrian move continuously. To evaluate our throughput-prediction model and collect the learning data required for training the DNN, we developed an actual indoor experimental setup where 5G throughput and physical-space information are automatically measured using autonomous humanoid robots. The throughputs, including the sharp drops due to self-blocking by UE rotation and the blocking by an object moving in front of the UE, were captured. We showed that our model was effective in using surrounding object information as well as UE information for predicting 5G throughput one second ahead. Our model with physical-space information improved prediction accuracy by 57.5% at the 50th-percentile absolute error value compared with a prediction model that uses only the past throughput as the input feature. We will continue this research to develop core technologies toward 6G/IOWN.

For the further details of this article, please refer [11].

REFERENCE

[1] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, “Business Case and Technology Analysis for 5G Low Latency Applications,” IEEE Access, Vol. 5, pp. 5917–5935, Apr. 2017.
https://doi.org/10.1109/ACCESS.2017.2685687

[2] NTT DOCOMO, “DOCOMO 6G White Paper,”
https://www.docomo.ne.jp/english/corporate/technology/whitepaper_6g/

[3] IOWN,
https://www.rd.ntt/e/iown/

[4] T. Cogalan and H. Haas, “Why Would 5G Need Optical Wireless Communications?”, Proc. of the 28th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC 2017), pp. 1–6, Montreal, Canada, Oct. 2017.

[5] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, “Millimeter-Wave Communications: Physical Channel Models, Design Considerations, Antenna Constructions, and Link-Budget,” IEEE Commun. Surveys Tuts., Vol. 20, No. 2, pp. 870–913, 2nd quarter 2018.

[6] J. Mashino, Y. Fujino, and R. Kudo, “R&D Activities of Core Wireless Technologies toward 6G Radio Access,” NTT Technical Review, Vol. 20, No.7, pp. 30–36, July 2022.
https://doi.org/10.53829/ntr202207fa4

[7] T. Nishio, H. Okamoto, K. Nakashima, Y. Koda, K. Yamamoto, M. Morikura, Y. Asai, and R. Miyatake, “Proactive Received Power Prediction Using Machine Learning and Depth Images for mmWave Networks,” IEEE J. Sel. Areas Commun., Vol. 37, No. 11, pp. 2413–2427, Nov. 2019.
https://doi.org/10.1109/JSAC.2019.2933763

[8] H. Nagata, R. Kudo, K. Takahashi, T. Fujita, K. Takasugi, Y. Aoki, Y. Horise, and Y. Morihiro, “5G Throughput Prediction for 28 GHz Channels Using Physical Space Information,” Proc. of the 2024 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, Dubai, United Arab Emirates, Apr. 2024.
https://doi.org/10.1109/WCNC57260.2024.10571119

[9] ROS,
http://wiki.ros.org/navigation

[10] Iperf,
https://iperf.fr/

[11] H. Nagata, R. Kudo, K. Takahashi, T. Yamada, T. Fujita, “5G Throughput-prediction Technology for 28-GHz Channels Using Physical-space Information,” NTT Technical Review, Vol. 23, No.3, pp. 42–48, March 2025.
https://doi.org/10.53829/ntr202503ra1

II-14. 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.**

II-14.1. 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].

II-14.2. DCNN-based Radio Propagation Prediction Model

II-14.2.1. DCNN Configuration

DCNN of our proposed model is constructed by two parts: feature extraction part and prediction part, as show in Fig. II-14.2.1-1.

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. $f(x) = \max(0, x)$.

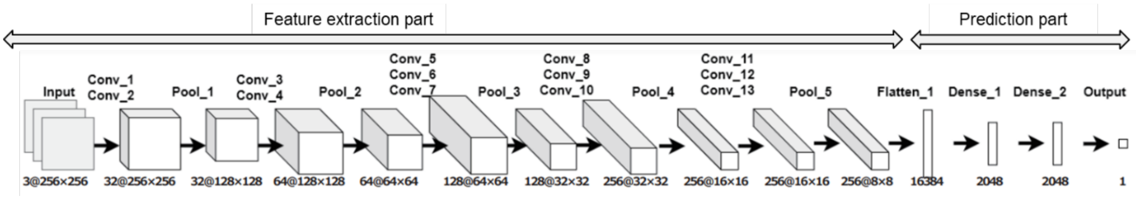


Fig. II-14.2.1-1 DCNN configuration

II-14.2.2. 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. II-14.2.2-1, 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.

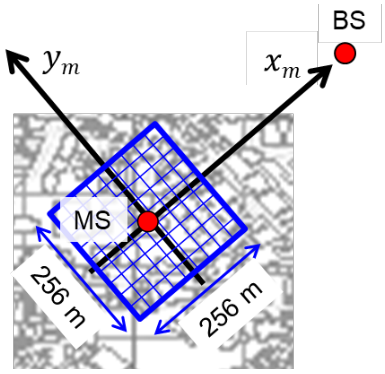
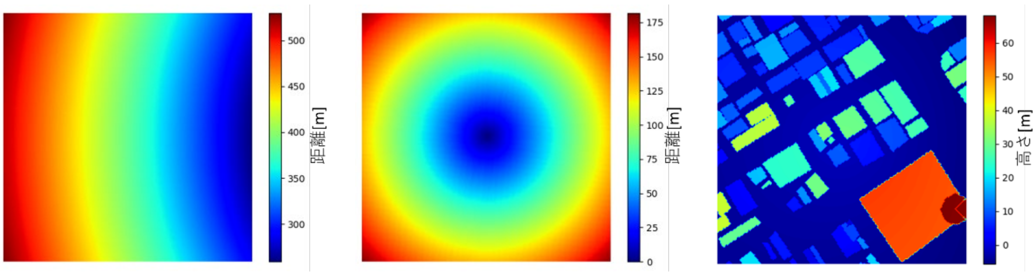


Fig. II-14.2.2-1. Definition of rectangular region

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 II-14.2.2-2 shows the examples of input map data.



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

Fig. II-14.2.2-1. Examples of input map data

II-14.3. Performance of DCNN-based Model

II-14.3.1. 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 II-14.3.1-1 and Table II-14.3.1-1 show the measurement area and conditions, respectively.

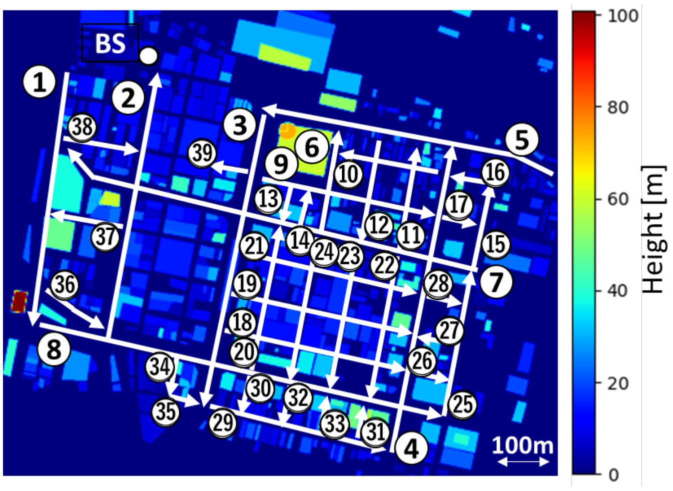


Fig. II-14.3.1-1 Measurement area (Kyushu Kokura area, Japan): White lines represent measurement courses.

Table II-14.3.1-1 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.

II-14.3.2. Evaluation Results

Figure II-14.3.2-1 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.

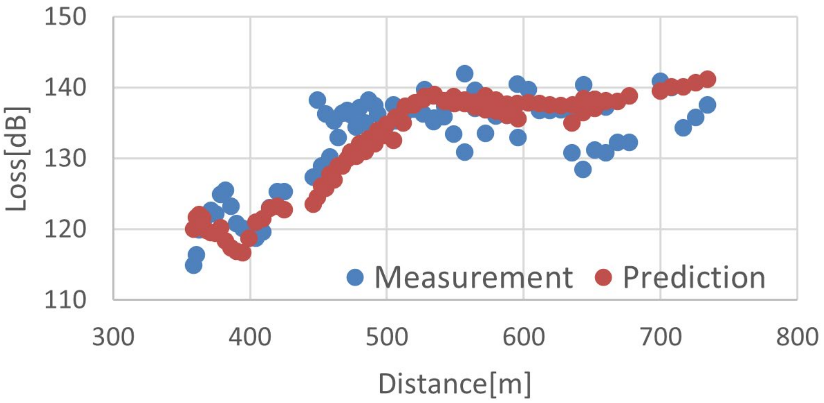
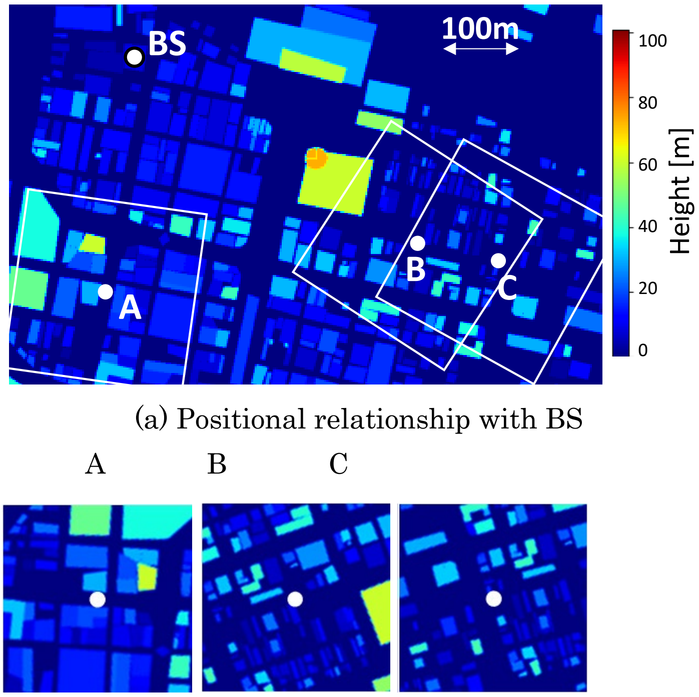


Fig. II-14.3.2-1 Prediction results.

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. II-14.3.2-2. Figure II-14.3.2-3 shows the analysis results with Grad-CAM. In Fig. II-14.3.2-3, 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. II-14.3.2-1 Reception points for evaluation of extracted features from map data.

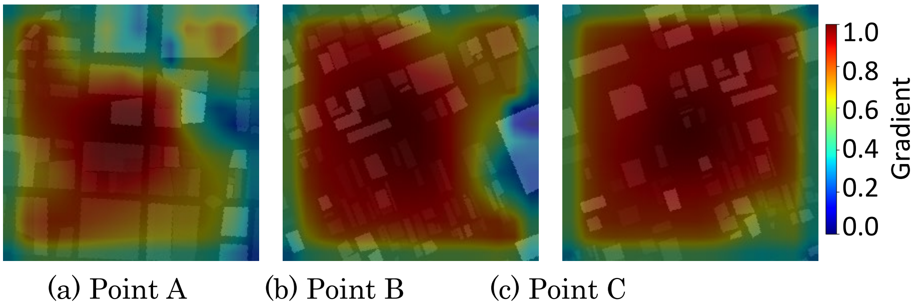


Fig. II-14.3.2-2 Analysis results with Grad-CAM when using multiple maps.

Finally, Fig. II-14.3.2-3 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 II-14.3.1-1. 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.

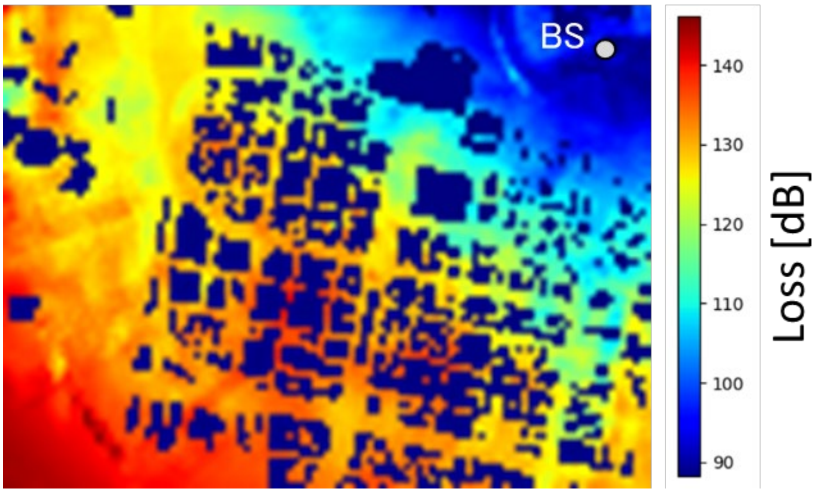


Fig. II-14.3.2-3 Propagation loss distribution predicted by trained DCNN.

II-14.4. 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. 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.

II-15. AI-Based Radio Propagation Modeling for Wireless Emulator

Tatsuya Nagao, Takahiro Hayashi
KDDI Research, Inc.

***Abstract*— To efficiently design and validate wireless communication systems, including Beyond 5G, research and development of wireless emulators that replicate the behavior of wireless communications in a virtual environment is progressing. Precise emulation requires accurate models of radio propagation characteristics in real environments. We introduce recent advancements in site-specific radio propagation modeling techniques utilizing machine learning.**

II-15.1. Introduction

In the design of wireless communication systems, the verification and performance evaluation of systems based on real-world use cases are critical processes. However, conducting field tests using actual wireless devices in real environments requires substantial resources and poses challenges in ensuring reproducibility. Consequently, research and development efforts are advancing toward wireless emulators replicating communication environments in a virtual space, thereby simulating the behavior of wireless communication systems [1]. These wireless emulators aim to construct a digital twin of wireless communication by enabling wireless communication systems, composed of virtual devices built in a virtual space and those connected via physical interfaces, to operate in real-time to simulate the system's dynamic characteristics.

When evaluating and validating wireless communication systems using wireless emulators, it is desirable that the radio propagation characteristics in the expected usage environment can also be reproduced in the virtual space. Traditional radio propagation models are typically constructed based on statistical processing of simple environmental parameters, such as the distance between Tx and Rx and measured data. However, as actual propagation characteristics can vary significantly due to surrounding environments, the accuracy of site-specific propagation characteristics proves insufficient for precise emulation of wireless communication.

To address this, various methods utilizing machine learning to establish models that consider site-specific environmental information are being investigated. By implicitly learning the relationships between environmental data and measured data, models tailored to individual locations can be constructed. Additionally, the application of artificial intelligence (AI) techniques, such as image recognition, facilitates the extraction of features from multidimensional data like environmental spatial information, thereby enabling complex pattern recognition.

II-15.2. Path Loss Model Based on Residual Networks (ResNet)

This section describes the proposed method for modeling path loss [2]. As illustrated in Fig. II-15.2-1, the method utilizes three types of map data as input: (a) relative building height surrounding the Tx, (b) relative building height surrounding the Rx, and (c) the distance between the Tx and Rx. Here, relative building height refers to the height of a building relative to the height of the antenna, serving as an indicator of line-of-sight from the antenna. The extraction of map data is conducted to ensure that the directions from the Tx point to the Rx point, and vice versa, are aligned.

Furthermore, as shown in Fig. II-15.2-2, we have designed an architecture suitable for path loss prediction based on Residual Networks (ResNet), which are widely used in image recognition tasks. ResNet incorporates shortcut connections between several convolutional layers, allowing for efficient propagation of error information from the output layer back to the higher layers—specifically, to those layers closer to the input—during the training process. This structural characteristic enhances the model's ability to learn complex representations and improves the accuracy of path loss predictions in the presence of varying environmental factors.

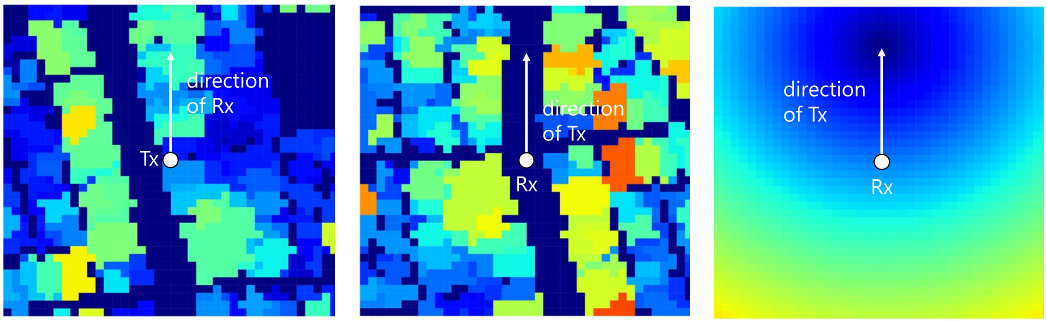


Fig. II-15.2-1. Examples of Input Data

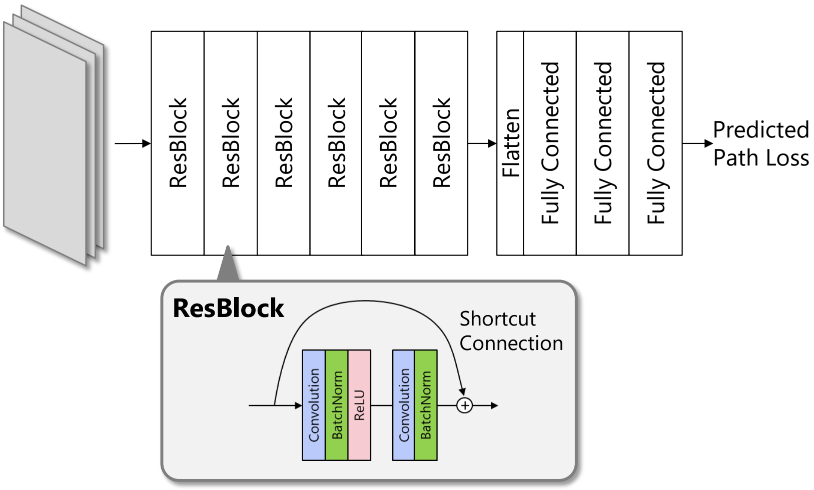


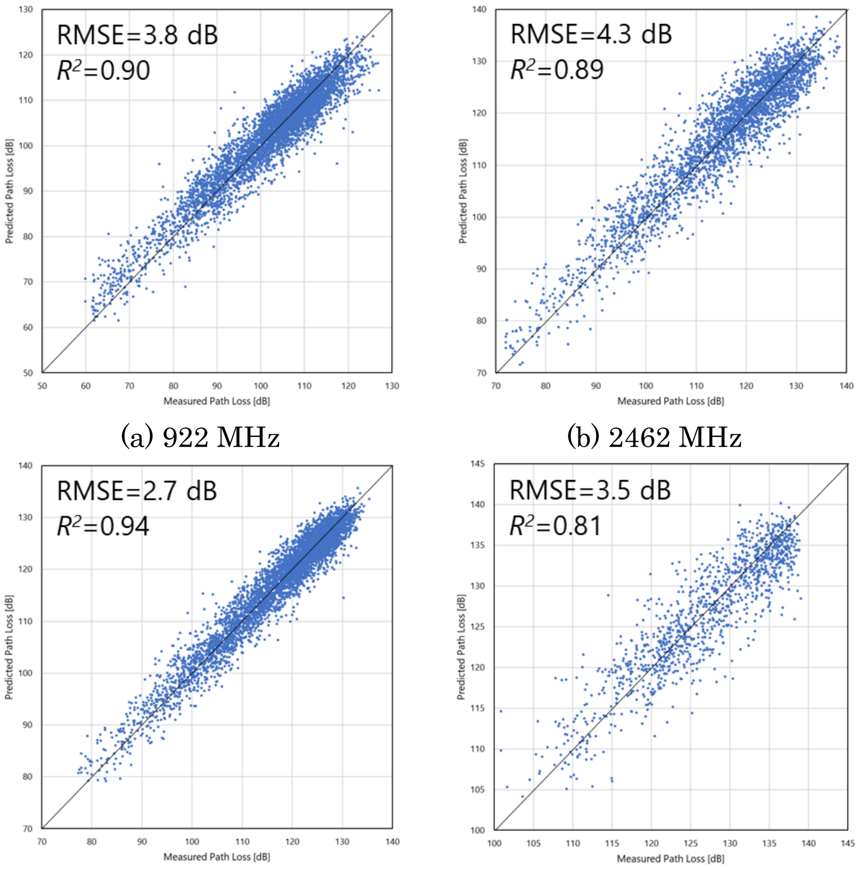
Fig. II-15.2-2. Path Loss Model Based on Residual Networks

II-15.3. Evaluation Results

The accuracy of path loss prediction using the proposed method was evaluated based on measured data obtained from an urban area in Yokohama City. A Tx was set up on the rooftop of a building approximately thirty-one meters high, where measurements were conducted across four frequency bands: 922 MHz, 2462 MHz, 4850 MHz, and 28.35 GHz [3]. K-folding Cross-validation (K=5) was employed for the assessment of the proposed method. Specifically, 80% of the dataset was used for training, while the remaining 20% was reserved for evaluation, with this process repeated five times. A comparative evaluation was also conducted against the widely recognized statistical model, the 3GPP TR 38.901 Urban Macro (UMa) model [4]. The evaluation results are presented in Table. II-15.3-1 and Fig. II-15.3-1. As seen from the table, the proposed method significantly improves prediction accuracy compared to the UMa model. Furthermore, the predicted values from the proposed method closely correspond to the measured values, as illustrated in the figures.

Table. II-15.3-1. Evaluation Results

Frequency	RMSE [dB]	
	3GPP Urban Macro (UMa)	ResNet (proposed)
922 MHz	8.8	3.8
2462 MHz	7.4	4.3
4850 MHz	8.2	2.7
28.35 GHz	17.6	3.5



(a) 922 MHz (b) 2462 MHz

(c) 4850 MHz (d) 28.35 GHz

Fig. II-15.3-1. Scatter Plot
(x: Measured Path Loss, y: Predicted Path Loss)

II-15.4. Conclusion

This article introduces methodologies for applying AI technologies to radio propagation modeling, a critical component for realizing wireless emulators as digital twins of wireless communication systems. By utilizing site-specific environmental information through machine learning, we have demonstrated the ability to simulate site-specific radio propagation characteristics accurately. The findings of this study are expected to contribute to the efficient design and optimization of future wireless communication systems, representing a significant step forward in the evolution of wireless communication technology.

Acknowledgements

This research has been conducted under the contract "R&D on the deployment of mobile communication systems in the millimeter wave band and other frequencies" (JPJ000254) made with the Ministry of Internal Affairs and Communications of Japan.

REFERENCE

[1] F. Kojima, T. Miyachi, T. Matsumura, H. Sawada, H. Harai and H. Harada, "A Large-Scale Wireless Emulation Environment with Interaction between Physical and Virtual Radio Nodes for Beyond 5G Systems," 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 2022.

[2] T. Nagao and T. Hayashi, "A Study on Path Loss Modeling using ResNet and Pre-Training with Free Space Path Loss," 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 2022.

[3] T. Kumagai, et al., "A Study on Propagation Loss Measurement Scheme using CW Signal under Burst-Interference Environments," AP2022-238, 2023. (Japanese)

[4] 3GPP TR 38.901 (V17.0.0): Study on channel model for frequencies from 0.5 to 100 GHz (Release 17), 2022.

[5] T. Nagao, “AI-based Radio Propagation Modeling and Challenges,” IEICE Society Conference, TK-3-02, Sept. 2024.

II-16. 6G Simulator Utilizing Future Prediction Control Technology Based on AI/ML

Yuyuan Chang, Yuta Hayashi, Kiichi Tateishi, Satoshi Suyama, and Huiling Jiang
NTT DOCOMO, INC.

II-16.1. Introduction

In 6G, to achieve communication speeds exceeding 100 Gbps, the utilization of frequency bands called sub-terahertz bands, such as 100 GHz, which can use greater bandwidths, is being considered compared to 5G [1]. Additionally, discussions are beginning on the use of frequency bands known as mid-bands, ranging from 7 GHz to 24 GHz. Similar to 5G, it is anticipated that communication systems will be constructed by combining two types of frequency bands. To realize ultra-low latency communication, high connectivity, and coverage assurance, a distributed network enhancement technology (NRNT: New Radio Network Topology) is being proposed [2], which will establish a distributed network topology in the spatial domain. For example, new network forms such as reconfigurable intelligent surfaces (RIS) that can control reflection directions and intensities, and moving base stations (BS) like base station drones are being envisioned.

In addition to the advancement of conventional wireless communication technologies, the utilization of AI (Artificial Intelligence) is also being considered. In the 6G era, it is anticipated that vast and diverse information such as images, audio, and video will be transmitted from various devices, and AI technology is expected to be used to analyze and leverage this extensive and varied information. Furthermore, the introduction of AI technology into wireless communication systems is being contemplated, with the expectation that it will provide higher quality communication by implementing various controls in wireless communication, managing networks and devices, and automating optimization functions for use cases and environments. Particularly in the fusion of cyber-physical spaces, video and various sensing information will be transmitted to the network through IoT (Internet of Things) devices. Based on the transmitted information, calculations can be performed in cyber space to predict a few seconds ahead, and the predicted information can be utilized in the physical space for precise communication, such as base station selection and beam selection.

The authors have developed a 6G system-level simulator (6G simulator) designed to evaluate and visualize the technologies being considered for 6G as a whole system [3]-[5]. Figure II-16.1-1 illustrates the worldview of the 6G simulator. So far, the sub-terahertz band, mid-band, and NRNT have been integrated into the 6G simulator, and evaluations have been conducted in a virtual outdoor urban environment. Additionally, machine learning (ML) algorithms from AI technology have been incorporated into the 6G simulator, enabling predictive control to avoid the impact of unexpected obstacles

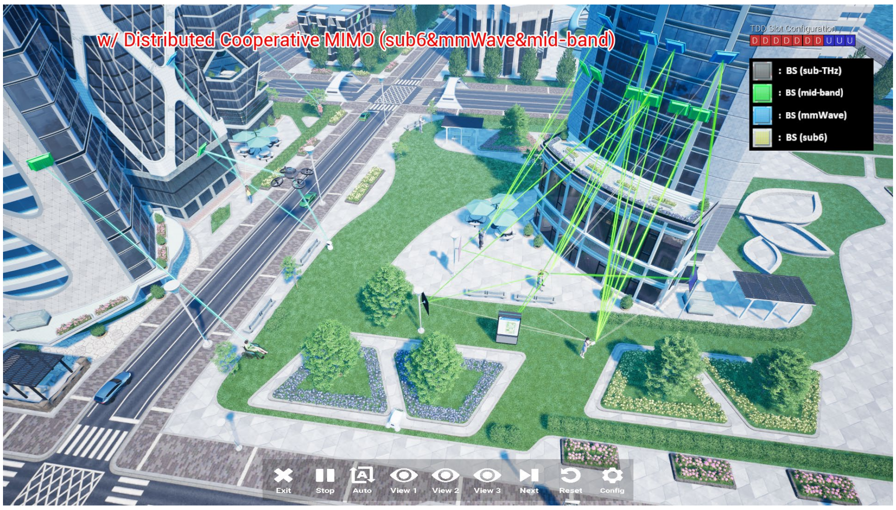


Figure II-16.1-1 The worldview of the 6G simulator.

based on pre-learned results, demonstrating the use cases of AI technology in wireless communication systems [6]. This work investigates the effects of using AI-ML under various conditions on system performance, thereby clarifying the effectiveness of wireless communication systems utilizing AI technology. Simulations will also be conducted under various conditions, such as changes in the TRP (Transmission and Reception Point) selection cycle and the mobility speed of user equipment (UE), to confirm how the effects of AI-ML manifest and to consider effective use cases for AI-ML [6], [7].

II-16.2. Future Prediction Control Using AI-ML Technology

In 5G and 6G, high-frequency bands are used, which have strong directivity and high propagation loss, however, enabling high-speed, large-capacity communication and low-latency communication. Therefore, it is important to create environments that are as close and unobstructed as possible, and the use of RIS and BS drones to intentionally establish communication pathways is being considered. However, since the communication environment is constantly changing, there is a high possibility of throughput degradation due to the emergence of unexpected obstacles. Hence, it is conceivable to use future prediction control technology to avoid the impact of obstructions and prevent throughput degradation.

With the extreme-high-speed, extreme-large-capacity, and extreme-low latency communication features of 6G, the realization of autonomous driving for vehicles utilizing cellular networks is anticipated in the 6G era. In this work, we consider a situation where an autonomous vehicle, as shown in Figure II-16.2-1, is driving in an outdoor urban area, and the communication between the autonomous vehicle and the TRP located along the road involves the occurrence of obstructions. At that time, future

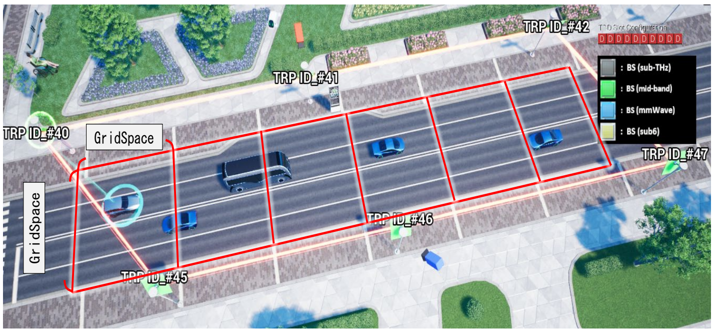
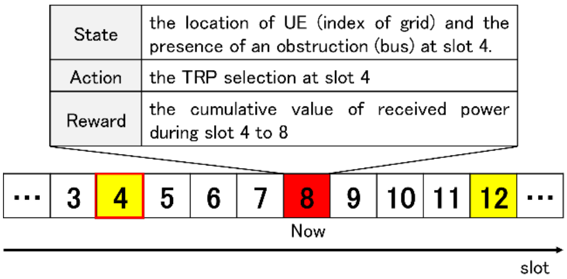


Figure II-16.2-1 The environment
(The image of the grid is shown by red lines)

prediction control using AI-ML will be implemented to prevent throughput degradation and ensure that a high throughput is consistently maintained.

The real communication environment is constantly changing, influenced by various factors such as time of day, weather, population density in the area, and the presence or absence of obstructions. Therefore, in this simulation, deep reinforcement learning is employed for TRP selection. Deep reinforcement learning is a technology that combines deep learning and reinforcement learning, enhancing decision-making capabilities in more complex environments by integrating the two approaches.

To implement deep reinforcement learning, it is necessary to define the environment, state, action, and reward. In this simulation, the environment is defined as "communication between the TRP on a straight road and the autonomous vehicle," and the state is defined as "the UE's location information and the presence of a bus that acts as an obstruction." The UE's location information is represented by dividing the AI-ML application area into a grid pattern, as shown in Figure II-16.2-1, using the grid numbers. Additionally, the length of one side of the grid is defined as "GridSpace." The action is defined as "the TRP selection process," and the reward is defined as "the cumulative value of received power within the TRP selection cycle." Furthermore, the ε-greedy policy is employed as the action policy. The ε-greedy policy allows for a balanced combination of "exploration," where actions are selected randomly by varying ε, and "exploitation," where actions are chosen based on rewards obtained from previous explorations, resulting in an action policy that is more suitable for the environment [7].



State	the location of UE (index of grid) and the presence of an obstruction (bus) at slot 4.
Action	the TRP selection at slot 4
Reward	the cumulative value of received power during slot 4 to 8

Figure II-16.2-2 An example of updating the neural network.
(Updating by every 4 slots)

Table II-16.3-1 AI-ML learning parameters

Learning method	Deep reinforcement learning (ε-greedy policy + Adam)
Discount factor	0.5
Learning rate	0.003
Size of mini-batch	64
Minimum number of episodes	1000
GridSpace	20 m
ε	0.001 ≤ ε ≤ 1.0

Table II-16.3-2 simulation parameters

Center frequency	15 GHz
Band width	400 MHz
No. of antennas of TRP	9
The Tx power of TRP	30 dBm
Height of TRP	10 m
The period of TRP selection	10, 20, 40 ms
No. of antennas of UE	144 (4V × 4H × 9 panels)
The Tx power of UE	23 dBm
Height of UE	1.5 m
The speed of UE	30, 60, 120 km/h

For constructing the neural network, Adam (Adaptive Moment Estimation) [8] is used. The results obtained from the ε-greedy policy are utilized to update the neural network. The update timing is aligned with the TRP selection cycle. Figure II-16.2-2 illustrates an example of updating the neural network. For example, let's assume the TRP selection cycle occurs every 4 slots. If the current time is slot 8, the UE's location and the presence of a bus at slot 4, as well as which TRP was selected, will be learned based on the received power obtained from the connected TRP up to slot 8. By performing this process for each TRP selection cycle, it becomes possible to select the TRP that yields the highest received power within the selection cycle, taking into account the presence or absence of obstructions.

II-16.3. Simulation Using a 6G Simulator

Table II-16.3-1 shows the learning parameters for AI-ML, and Table II-16.3-2 presents the simulation parameters. The flow of the simulation begins with the generation of the learning model. During this process, the parameter values indicated in Table II-16.3-1 are used for training. Subsequently, the generated learning model is employed for predictive control. In this simulation, the frequency range of 15 GHz in the mid-band is utilized. Each TRP consists of nine antennas, while the UE is configured with nine 4×4 subarrays. Each UE forms beams through hybrid beamforming. The UE selects the TRP and beams that provides the highest received power and connects to the TRP with that beam. Furthermore, the number of TRPs installed is set to nine, and one UE is chosen for evaluation. In this simulation, data is transmitted using time division duplex (TDD) with a downlink to uplink ratio of 10:0. Therefore, this report focuses on evaluating the downlink throughput. The blockage caused by a bus utilizes a model based on TR38.901's Blockage model B [9]. This model arranges flat obstructions and applies attenuation based on the difference between the straight-line distance between the transmission and reception points and the distance via the top, bottom, left, and right edges of the obstruction. In this report, the movement speed of UE is categorized into three types: 30 km/h, 60 km/h, and 120 km/h. In the 6G simulator, at 30 km/h, this corresponds to moving 1 meter per slot; at 60 km/h, it corresponds to 2 meters, and at 120 km/h, it corresponds to 4 meters.

II-16.3.1. Effects of AI-ML under Different TRP Selection Cycles

Figures II-16.3.1-1 to II-16.3.1-3 show the variation in throughput when the mobile station (MS) moves at a speed of 60 km/h, with a GridSpace of 20 m, and TRP selection cycles of 10, 20, and 40 ms. The shaded areas in the figures indicate the timing of bus stops. The solid lines represent the characteristics when AI-ML is applied, while the dashed lines indicate the

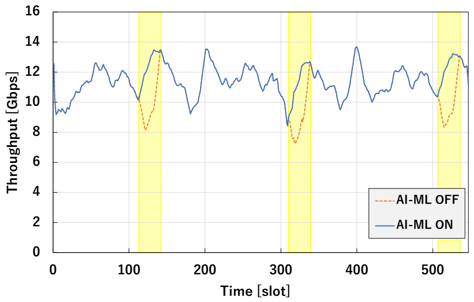


Figure II-16.3.1-1 variation in throughput (selection cycle: 10 ms)

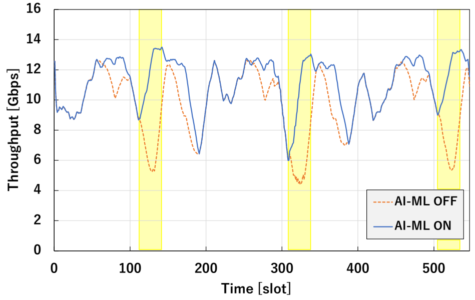


Figure II-16.3.1-2 variation in throughput (selection cycle: 20 ms)

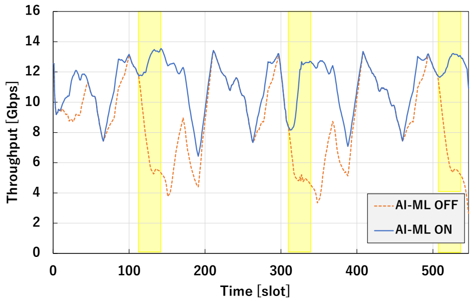


Figure II-16.3.1-3 variation in throughput
(selection cycle: 40 ms)

characteristics without the application of AI-ML. Focusing on the shaded areas, it can be seen that in all patterns, the use of AI-ML helps prevent a decline in throughput. Moreover, as the TRP cycle increases, the throughput tends to improve regardless of the presence or absence of obstructions, and the effect of AI-ML becomes more pronounced. When the TRP selection cycle is long, the connection remains with a single TRP, allowing for the selection of one with high received power during the connection moment. However, considering the overall received power, it is more likely to be lower. Conversely, by using AI-ML, it becomes possible to select a TRP that will increase the received power from the current selection to the next, which suggests, as shown in Figures II-16.3.1-2 and II-16.3.1-3, that the effects of AI-ML manifest strongly even when obstructions do not appear. When the TRP selection cycle is short, the throughput characteristics do not change regardless of the application of AI-ML when there are no obstructions. This is because, with a short TRP selection cycle, it is possible to continuously select TRPs with high received power, resulting in good throughput characteristics without the need to consider maximizing received power within the selection cycle. However, considering real-world communication systems, processing delays may occur, making rapid selection difficult. Therefore, it is expected that in realistic environments, the characteristics will resemble those with a longer TRP selection cycle as shown in Figures II-16.3.1-2 and II-16.3.1-3, making the introduction of AI-ML effective. It should be noted that there are moments when throughput significantly deteriorates regardless of whether AI-ML is applied; this is due to processing in the simulator. As a result, similar downturns will occur in subsequent results.

II-16.3.2. Effects of AI-ML at Different UE Speeds

Figures II-16.3.2-1 and II-16.3.2-2 show the results when the GridSpace is set to 20 m, the TRP selection cycle is 10 ms, and the speed of UE varies between 30 km/h and 120 km/h. The differing number of shaded areas in the figures (indicating the occurrence of obstructions) corresponds to the change in the speed of the bus that cause the obstructions, which is adjusted according to the UE's speed, resulting in variations in the number of times the bus travels along the measured road during the simulation time. Upon examining the throughput characteristics, when the UE's speed is 30 km/h, the use of AI-ML allows for the avoidance of obstruction effects, leading to improvements in throughput. In contrast, at 120 km/h, the effect of AI-ML is minimal. The throughput characteristics without the application of AI-ML do not deteriorate even when obstructions occur, suggesting that in this simulation, the configurations for the TRP selection cycle and TRP placement ensure that TRPs on the obstruction side are not selected even without AI-ML. Consequently, since the selected TRP remains unchanged with the application of AI-ML, there is no change in the throughput characteristics at the moments when obstructions occur. Figure II-16.3.2-2 presents the throughput

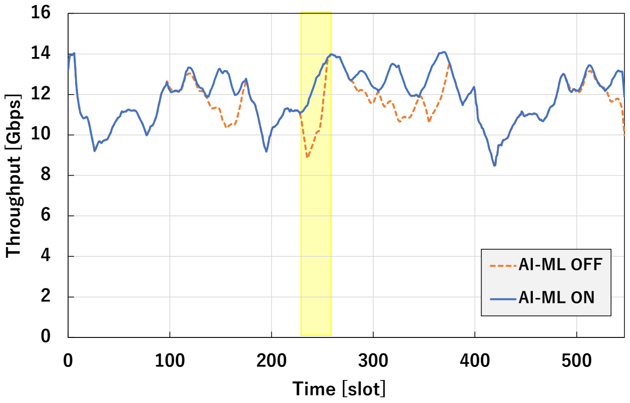


Figure II-16.3.2-1 variation in throughput
(speed: 30 km/h)

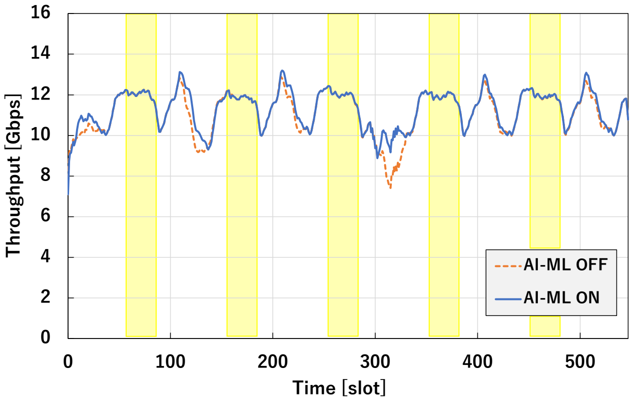


Figure II-16.3.2-2 variation in throughput
(speed: 120 km/h)

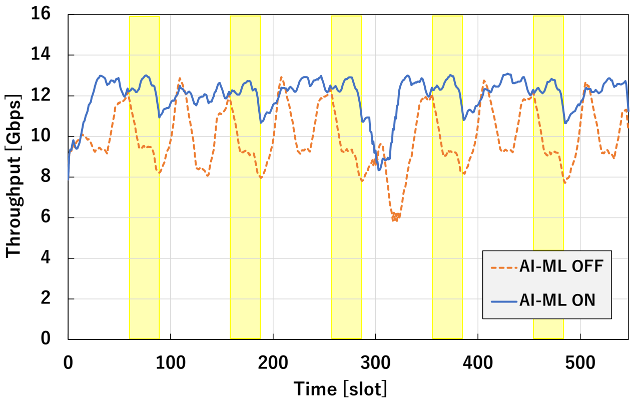


Figure II-16.3.2-3 variation in throughput
(speed: 120 km/h, selection cycle: 20 ms)

characteristics when the UE's speed is 120 km/h and the TRP selection cycle is set to 20 ms. The lengthened TRP selection cycle resulted in significant differences in characteristics depending on the application of AI-ML compared to the results shown in Figure II-16.3.2-2. By observing the occurrence of obstructions, it can be seen that the change in the TRP selection cycle has made the system susceptible to obstruction effects even at 120 km/h. Additionally, the use of AI-ML effectively mitigates the throughput degradation caused by these obstructions. As mentioned in the previous section, when the TRP selection cycle increases, even if the received power is high at the moment of selection, the overall received power within the cycle is likely to be low. Moreover, with higher speeds, the UE travels longer distance before the next TRP selection, significantly increasing the distance between the TRP and the UE. Thus, by selecting TRPs to maximize received power within the cycle via AI-ML, differences in throughput are observed depending on whether AI-ML is applied, as illustrated in Figure II-16.3.2-3.

II-16.4. Conclusions

In the simulations, AI-ML was implemented for communication between a single autonomous vehicle and a TRP installed along a straight road in an outdoor urban environment. By maximizing the received power within the selection cycle, it is possible to prevent throughput degradation caused by obstruction effects; thus, using AI-ML for predictive control in communications between vehicles and TRPs can be considered effective. Additionally, even when there are no obstructions, the impact of maximizing received power within the selection cycle is evident, and throughput significantly improved in environments where UE cannot consecutively select TRPs. This indicates that AI-ML can serve as a means to mitigate throughput degradation caused by processing delays or other system performance issues encountered during connections with TRPs. However, it should be noted that this simulation considers only one UE, and does not account for interference from other UEs. Furthermore, since the road is straight and the UE's movement path follows a single pattern, the environment is relatively easy to learn. When evaluating under real-world conditions, it is necessary to consider scenarios with an increased number of UEs or irregular movement of UEs. As a future prospect, we are considering the exploration of additional use cases for AI technology by examining new scenarios and adding learning parameters.

REFERENCE

[1] NTT DOCOMO, Inc., 5G Evolution and 6G White Paper (Version 5.0), Jan. 2023. https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v5.0.pdf

[2] M. Iwabuchi, S. Suyama, T. Arai, M. Nakamura, K. Goto, R. Ohmiya, D. Uchida, T. Yamada, T. Ogawa, “Concept and Issues of New Radio Network Topology for 5G Evolution & 6G,” IEICE Tech. Report, RCS2022-148, Oct. 2022.

[3] T. Okuyama, S. Suyama, N. Nonaka, T. Asai, “6G System-Level Simulator: Toward Realizing Extreme-High Date Rate Communication at 100 Gbps in the 100 GHz Band,” NTT DOCOMO Technical Journal, vol.29, no.3, pp.13-24, Oct. 2021.

[4] K. Tateishi, K. Kitao, S. Suyama, T. Yamada, “Advancement of 6G System-Level Simulator,” NTT DOCOMO Technical Journal, vol.31, no.2, Jul. 2023.

[5] K. Tateishi, S. Suyama, H. Jiang, “Real-Time Simulator for Sixth-Generation Mobile Communications System,” IEICE Tech. Report, RCS2023-143, Oct. 2023.

[6] K. Tateishi, Y. Hayashi, S. Suyama, Y. Chang, H. Jiang, “6G System Level Simulator for Evaluating Introduction of Mid-Band and Future Predictive Control Technology,” IEICE Tech. Report, SR-2024-75, Jan. 2025.

[7] Y. Hayashi, K. Tateishi, S. Suyama, Y. Chang, H. Jiang, “Evaluation of Future Prediction Technology Using Machine Learning by 6G System Level Simulator,” IEICE Tech. Report, RCS2024-264, Mar. 2025.

[8] D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” ICLR 2015, May 2015.

[9] 3GPP TR38.901 V16.1.0, Rel-16, 5G: Study on Channel Model for Frequencies from 0.5 to 100 GHz, Nov. 2020.

II-17. Optimization of 6G Radio Access Using Digital Twin

Yuyuan Chang, Koshiro Kitao, Takahiro Tomie, Nobuaki Kuno, and Satoshi Suyama
NTT DOCOMO, INC.

II-17.1. Introduction

In the advanced Cyber-Physical Systems (CPS) anticipated for the 2030s, artificial intelligence (AI) will recreate the real world in cyberspace through digital twins, allowing predictions of the future and new knowledge gained through emulation to be fed back into the real world, thereby providing various values and solutions. In this advanced CPS, to achieve high-capacity and low-latency transmission of sensing information from the real world to cyberspace, as well as reliable and low-latency control signal transmission from cyberspace to the real world, research and development of the 6G is being vigorously pursued [1]-[4]. In 6G, the peak data rate is set to exceed 100 Gbps, the area coverage rate providing Gbps-level services is 100%, the end-to-end (E2E) latency is less than 1 ms, the connection for an ultra-high number of devices is targeted at 10 million devices/km^2, and extremely stringent conditions such as ultra-low power consumption and low cost are established. Furthermore, 6G is expected to utilize AI technologies in every domain of the system, while wireless sensing technologies that leverage communication signal for sensing applications will enable high-precision terminal positioning with errors of less than a centimeter and surrounding object detection. For maximizing the performance of this 6G system, ensuring quality, and efficient system operations, dynamic control through CPS is expected to be introduced [5], [6].

Figure II-17.1-1 shows a 6G system utilizing dynamic control through CPS [6]. Here, the focus is primarily on the wireless portion of the 6G system. In the real space, there exists the actual 6G system, within which numerous wireless devices (base stations, relays, terminals, etc.) are connected to the wireless network. On the other hand, the virtual space contains propagation emulators, transmission emulators, and dynamic control. In the operational flow, first, at a certain point in time, a large volume of sensing information from the actual 6G system is sent in real-time to the virtual space. Based on this information, models of the usage environment and the wireless devices are constructed in the virtual space. Next, the propagation characteristics of each wireless

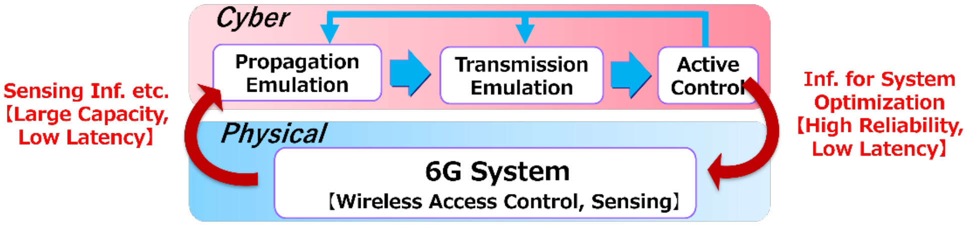

Cyber
Propagation Emulation
Transmission Emulation
Active Control
Sensing Inf. etc. [Large Capacity, Low Latency]
Inf. for System Optimization [High Reliability, Low Latency]
Physical
6G System
[Wireless Access Control, Sensing]

Figure II-17.1-1 a 6G system utilizing dynamic control through CPS.

link are estimated in succession through the propagation emulator and the transmission emulator, and the performance (throughput, interference level, power consumption, etc.) of each wireless device is evaluated based on the propagation characteristics. Furthermore, dynamic control optimizes the target performance in the virtual space. Using AI, it extracts information about the adjustment parameters and feeds this back to the propagation emulator and the transmission emulator. By repeatedly cycling through the propagation emulator, transmission emulator, and dynamic control, optimization can be achieved in the virtual space. Finally, as control signals necessary for the control of the actual 6G system, these are fed back to the real space and reflected in the actual 6G system. By enabling real-time cycles of high-capacity information exchange between the real space and the virtual space, it is believed that the maximum performance and efficient operation of the 6G system can be realized [7].

On the other hand, when utilizing high-frequency bands such as the sub-terahertz band in the system, it is necessary to not only validate individual technologies but also to conduct an early assessment of system performance when deploying multiple base stations (BS) and mobile stations (MS). This will clarify the performance improvement effects of the system as a whole and identify potential issues. However, device development generally requires significant time and cost, and it is necessary to ensure flexibility in changing configurations and parameters. Therefore, the authors aimed to demonstrate the feasibility of achieving ultra-high-speed communication through the utilization of the sub-terahertz band by developing a 6G system-level simulator (hereinafter referred to as the 6G simulator) and advancing its performance verification [8].

The conventional 6G simulator has the capability to evaluate the throughput when utilizing the 100 GHz band in two types of indoor environments, simulated as a shopping mall and a factory, confirming that throughput exceeding 100 Gbps can be achieved in both scenarios. When considering the introduction of 5G and 6G communication systems in specific environments such as indoors or factories, it is essential to understand the system performance, such as throughput, in advance for the intended environment. Moreover, visualizing the system performance provides very beneficial information for exploring methods of implementing communication systems. However, the conventional 6G simulator could only evaluate throughput in the pre-prepared scenarios and environments mentioned above. To accurately calculate throughput, high-precision estimation of radio wave propagation characteristics in the intended communication system environment is necessary. Recently, ray tracing calculations using polygon models of structures generated from point cloud data acquired in assessment environments have gained attention as methods to estimate propagation characteristics with high precision [9]-[12]. Therefore, the authors developed an enhanced 6G simulator capable of evaluating and visualizing the throughput of 5G and 6G based on the propagation characteristics obtained from ray tracing calculations using real environment models derived from point cloud data for the purpose of evaluating system performance in real environments [13]. This report introduces an overview of the functions of this simulator and examples of its performance evaluation [14], [15].

II-17.2. Overview of a 6G Simulator Using Real Environment Models Based on Point Cloud Data

This simulator is based on the 6G simulator reported in [8]. Below, we describe the overview of the conventional 6G simulator and this simulator. The conventional 6G simulator was developed to quantitatively validate the requirements and technical concepts of 6G as described in the NTT Docomo 6G white paper [1], as well as to verify the potential of utilizing the sub-terahertz band as a system [8]. In this simulator, we also aimed to apply the sub-terahertz band to achieve extreme-high data rate communication exceeding 100 Gbps more reliably, under the constraint of maintaining BS antenna sizes comparable to those of sub-6 and millimeter waves, and transmission power equivalent to that of 5G. By utilizing the sub-terahertz band, it is possible to significantly increase the number of antenna elements (hereinafter referred to as "elements") in Massive MIMO antennas, which in turn provides high beamforming (BF) gain that can compensate for the considerable propagation losses associated with the sub-terahertz band.

In the conventional 6G simulator, a channel model standardized by 3GPP was used for the simulation of the channel between the BS and MS at the system level [16]. In contrast, this simulator utilizes the propagation characteristic information obtained from ray tracing calculations applied to indoor real environment models generated from point cloud data. Specifically, it uses information on propagation loss, angles of arrival of waves, and propagation delay computed by ray tracing.

Table II-17.2-1 parameters for the ray tracing

Items	Parameters
Tool for ray tracing	Wireless Insight
Center frequency	4.7, 28, 100 GHz
Heights of BS and MS antennas	2.0 m and1.5 m
Type of BS and MS antennas	Omni (0 dBi)
Calculation Algorithm	ray launching
Angle interval between rays	0.25°
No. of reflections	7
Material	concrete

Figure II-17.2-1 shows an image of the evaluation environment in this simulator. This simulator is capable of displaying an evaluation environment image using point cloud data obtained from any environment, and the figure is generated using point cloud data collected from a conference room. By utilizing image data acquired simultaneously with the point cloud data from a camera, the actual conference room is reproduced in color. In the simulator, both the base stations (BS) and mobile stations (MS) can be positioned at arbitrary locations, and here, the evaluation environment is shown with 2 BSs and 6 MSs. When placing multiple BSs, by inputting the ray tracing calculation results for the placement of MS holistically within the evaluation area for each BS into the tool, it is possible to evaluate the throughput of the MS at any arbitrary location. However, the system does not accommodate MS movement.

Additionally, this simulator can visually capture the relationship between propagation characteristics and throughput characteristics by displaying a color map of the propagation parameters. The calculation parameters for the ray tracing are shown in

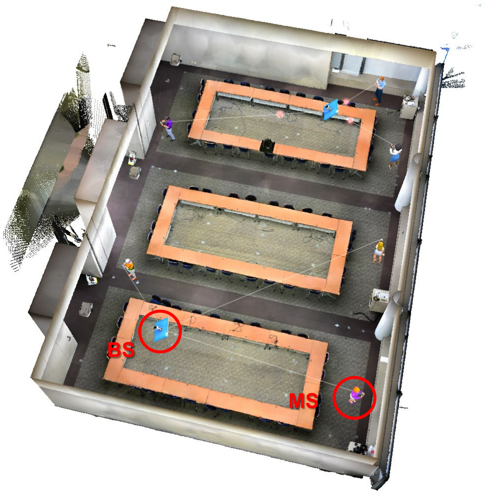


Figure II-17.2-1: An image of the evaluation environment in this simulator using point cloud data

Table II-17.2-1. In the ray tracing calculations, polygon data generated from point cloud data obtained in the conference room shown in Figure II-17.2-1 was input into Wireless Insight, a commercial ray tracing tool. The center frequencies were set to 4.7 GHz, 28 GHz, and 100 GHz, assuming configurations for 5G and 6G. The antennas for both the BS and MS are omnidirectional antennas, with the BS antenna height set at 2.0 m and the MS antenna height set at 1.5 m, and the ray search condition was set to 7 reflections. The material of the walls was calculated as concrete.

II-17.3. Evaluation Results of a 6G Simulator in a Conference Room Generated from Point Cloud Data

Examples of propagation characteristics calculated through ray tracing are shown in Figures II-17.3-1 to II-17.3-3. These figures represent color maps of the received level, delay spread, and angle spread in the horizontal plane on the MS side at 100 GHz. It can be observed that the received level is high near the BS, and due to reflections, the angle spread becomes larger near the walls of the conference room.

Next, based on the propagation parameters calculated from the above ray tracing, we describe the throughput characteristics computed by this simulator. The parameters for the system-level simulation using this simulator are shown in Table II-17.3-1. Fading channels are generated from the propagation loss, propagation delay, and angles of arrival calculated for each ray in the ray tracing, and the throughput is calculated when beamforming (BF) and MIMO spatial multiplexing are performed using Massive MIMO.

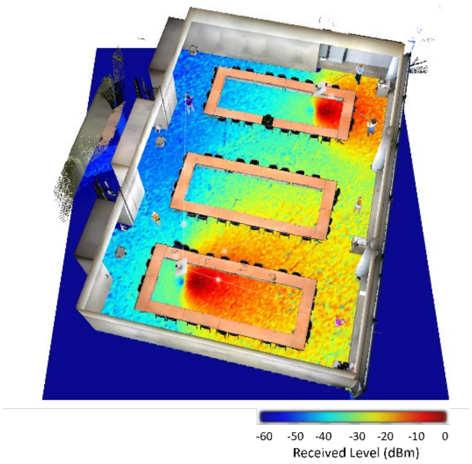


Figure II-17.3-1 the color map of received level

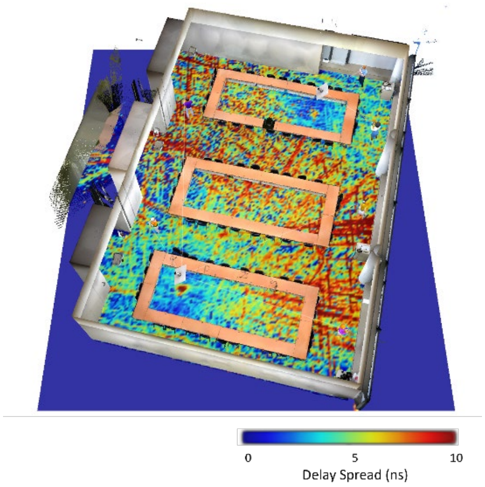


Figure II-17.3-2 the color map of delay spread

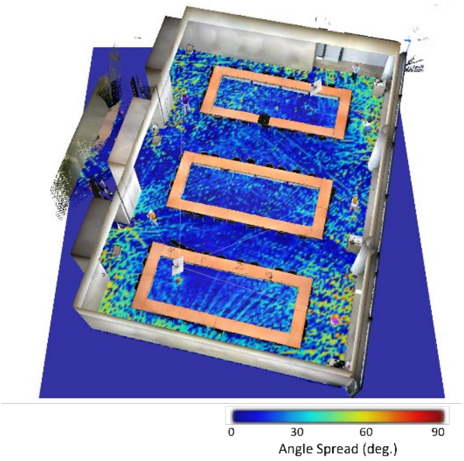


Figure II-17.3-3 the color map of angle spread

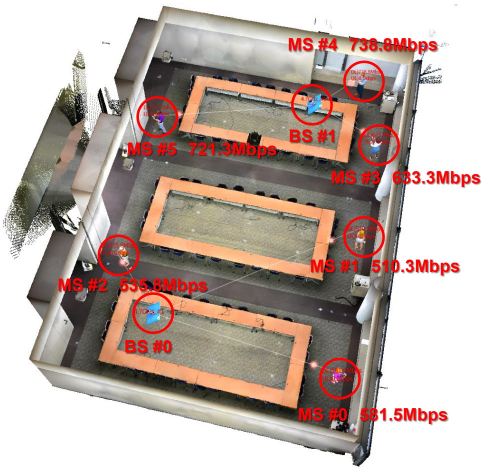


Figure II-17.3-4 the user throughput (4.7 GHz)

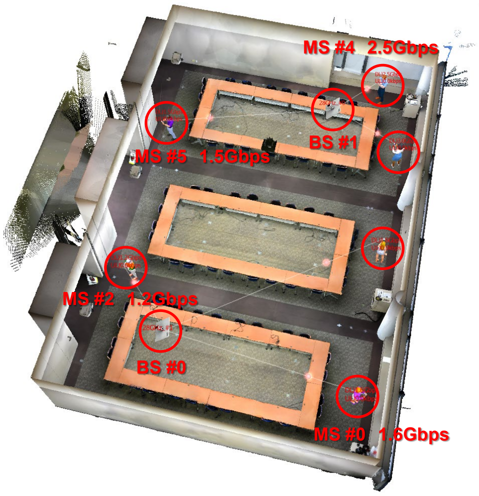


Figure II-17.3-5 the user throughput (28 GHz)

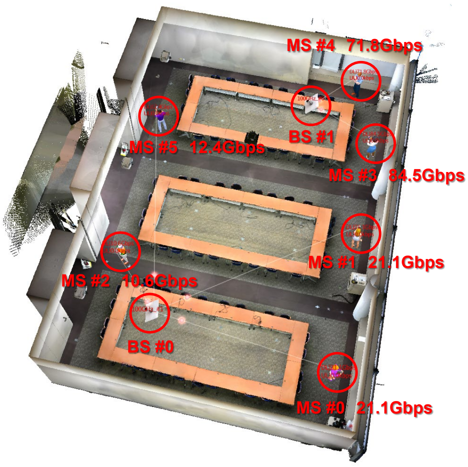


Figure II-17.3-6 the user throughput (100 GHz)

Table II-17.3-1 the parameters of system-level simulations

Center frequency	4.7 GHz	28 GHz	100 GHz
Band width	100 MHz	400 MHz	8.0 GHz
Tx power of BS	30 dBm		
Tx power of MS	23 dBm		
No. of BS antennas (V x H x sub-array)	144 (4 x 4 x 9)	2304 (16 x 16 x 9)	9216 (32 x 32 x 9)
Distance between sub-array	0.5 λ	4 λ	30 λ
No. of MS antennas (V x H x sub-array)	144 (4 x 4 x 9)		
Distance between MS elements	0.5 λ		
No. of BS	1		
No. of MS	2		
No. of MIMO layers	1, 2, 3, 4, 8		

Figures II-17.3-4 to II-17.3-6 show the downlink (DL) user throughput at 4.7 GHz, 28 GHz, and 100 GHz when two BSs and six users are placed. The figures display the user throughput for MS#0 to MS#5, indicated by red circles. It can be observed that MS#3 and MS#4, which are located in areas with high received levels and large angle spreads as shown in Figure II-17.3-1, achieve relatively high throughput. The average throughput of the six MSs at 4.7, 28, and 100 GHz is approximately 0.62, 1.9, and 37 Gbps, respectively. This confirms that utilizing higher frequency bands improves throughput due to the effects of increased bandwidth.

Furthermore, Figure II-17.3-7 shows the throughput map for the case of 100 GHz. Here, the BS is positioned as BS#0 in Figure II-17.3-6, and the throughput is displayed in a color map when only the position of one MS is changed. As in the previously mentioned cases, it can be seen that a throughput of 100 Gbps is achieved in areas with high received levels.

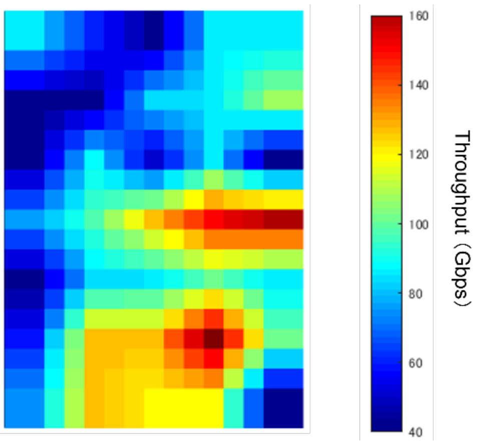


Figure II-17.3-7 the color map of user throughput (100 GHz)

II-17.4. Conclusions

In this work, we discussed the development of a simulator that can evaluate system performance in any environment using real environment models based on point cloud data as an enhancement of the 6G simulator. In the future, we plan to conduct performance evaluation and high-precision improvements of the simulator by comparing throughput measured using experimental equipment in real environments with the calculation results from this simulator. Additionally, towards the future development of tools that can dynamically control and optimize 6G using CPS, we will advance technical studies for high precision and fast processing in propagation simulations using real environment models, as well as link-level and system-level transmission simulations.

REFERENCE

[1] NTT DOCOMO, Inc., 5G Evolution and 6G White Paper (Version 5.0), Jan. 2023. https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v5.0.pdf

[2] Hexa-X, Deliverable D2.1, Towards Tbps Communications in 6G: Use Cases and Gap Analysis, Jun. 2021.

[3] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland and F. Tufvesson, "6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities," Proc. of the IEEE, vol. 109, no. 7, pp. 1166-1199, July 2021.

[4] T. S. Rappaport, *et. al.*, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," IEEE Access, June 2019.

[5] S. Suyama, K. Kitao, H. Tomie, K. Nakamura, T. Yamada, "Dynamic Control of 6G CPS Utilizing High-Speed Propagation Emulation," IEICE Society Conf., BS-1-11, Sept. 2022.

[6] H. Tomie, S. Suyama, K. Kitao, "Propagation Emulation Method for 6G Using Dynamic Control using CPS," IEICE Society Conf., BS-3-2, Sept. 2023.

[7] H. Tomie, S. Suyama, K. Kitao, N. Kuno, "Study on Propagation Emulation Method for 6G Using Dynamic Control by CPS," IEICE Tech. Report, RCS2023-170, Nov. 2023.

[8] T. Okuyama, S. Suyama, N. Nonaka, T. Asai, "6G System-Level Simulator: Toward Realizing Extreme-High Date Rate Communication at 100 Gbps in the 100 GHz Band," NTT DOCOMO Technical Journal, vol.29, no.3, pp.13-24, Oct. 2021.

[9] J. Jarvelainen, K. Haneda, and A. Karttunen, "Indoor Propagation Channel Simulations at 60 GHz using Point Cloud Data," IEEE Trans. on Antennas and Propagation, vol. 64, no. 10, pp. 4457-4467, Oct. 2016.

[10] P. Koivumaki, G. Steinbock, and K. Haneda, “Impacts of Point Cloud Modeling on the Accuracy of Ray-Based Multipath Propagation Simulations,” IEEE Trans. on Antennas and Propagation, vol. 69, no. 8, pp. 4737-4747, Aug. 2021.

[11] W. Okamura, R. Lukita, G. Ching, Y. Matsuyama, Y. Kishiki, Z. Chen, K. Saito, and J. Takada, “Simplification Method of 3D Point Cloud Data for Ray Trace Simulation in Indoor Environment,” IEICE Communications Express, vol. 9, no. 6, pp. 182-187, Jun. 2020.

[12] K. Kitao, M. Nakamura, T. Tomie, and S. Suyama, “Study of Raytracing using Point Cloud Data for Indoor Area Evaluation,” ICETC 2022, S1-6, Nov. 2022.

[13] K. Tateishi, K. Kitao, S. Suyama, T. Yamada, “Advancement of 6G System-Level Simulator,” NTT DOCOMO Technical Journal, vol.31, no.2, Jul. 2023.

[14] K. Kitao, S. Suyama, K. Tomie, N. Kuno, K. Tateishi, H. Jiang, “6G Simulator using Real Environment Model Generated from Point Cloud Data,” IEICE Tech. Report, AP-2023-65, July 2023.

[15] K. Kitao, S. Suyama, K. Tomie, N. Kuno, K. Tateishi, H. Jiang, “Enhancement of 5G/6G Simulator using Point Cloud Data,” IEICE Tech. Report, SR-2024-76, Jan. 2025.

[16] 3GPP TR38.901 V16.1.0, Rel-16, 5G: Study on Channel Model for Frequencies from 0.5 to 100 GHz, Nov. 2020.

II-18. Digital-Twin for and by Beyond 5G

Hideyuki Shimonishi, Osaka University
Kentaro Ishizu, Koji Zettsu, NICT
Toshirou Nakahira, Shoko Shinohara, NTT
Dai Kanetomo, NEC

***Abstract*— In this article, we discuss that a Digital-Twin can be both digital representation of real-world, i.e. Digital-Twin by Beyond 5G, which could be enabled by advanced Beyond 5G capabilities, and digital representation of network objects, i.e. Digital-Twin for Beyond 5G, which helps to enable advanced Beyond 5G capabilities. Federating and jointly optimizing various Digital-Twin instances of both real-world and network objects is essential to realize new services in the era of Beyond 5G, and thus propose a Digital-Twin architecture which manages various Digital-Twin instances in a common way so that any Digital-Twin applications can easily utilize them. We then introduce probabilistic Digital-Twin, which can improve both efficiency and safety of many Digital-Twin use cases by considering uncertainties inherent in the real world, and cross-domain orchestration of Digital-Twins, which will be a key to realize the digital-first services. Finally, we introduce some of examples of the Digital-Twins discussed in this article, including radio communication environment, human-robot cooperation, and smart sustainable mobility.**

II-18.1. Introduction

Digital-Twin is a digital reproduction of objects in physical space (cars, jet engines, people, buildings, cityscapes, etc.), or potentially in virtual space or so called Metaverse. It is expected to be an important technology for various ICT systems in factories, aviation, connected cars, smart cities, smart buildings, etc., in realizing advanced Cyber Physical Systems. The concept of Digital-Twin has been introduced in various literature since the 2000s, and widely accepted in recent years when several literatures, such as [1], has been known. Also, several articles, such as [2], provides an extensive survey on Digital-Twins, including enabling technologies and technical issues.

As an extension of classical notion of Digital-Twin, which is a one-to-one correspondence between objects in physical space and objects reproduced in virtual space, Digital Twin Network [3] has been proposed to represent of networks of multiple objects so that various objects in real and virtual space share information and cooperate to perform specific tasks in connected cars, smart cities, etc. In addition, Cognitive Digital Twin [4] focuses on knowledge representation, called ontology, to handle various types of objects in the real world. This is expected to enable more sophisticated applications by sharing digital twins between different systems.

Digital-twin, sometimes called Network Digital-Twin, is also used to plan, design, manage, simulate, operate, and control networks, as discussed by ITU-T [5], IOWN Global Forum Digital [6], and TMforum [7]. In this case, the Digital-Twin is a digital reproduction of any network devices, edge/cloud computing resources, terminal devices or robots which has network interfaces, radio communication environments measured or estimated by various sensing devices, or even logical networks and services.

II-18.2. Digital-Twin Platform Architecture with Beyond 5G

The key concept of the proposed Digital-Twin platform architecture is: 1) Digital-Twin instances of both real-world and network objects can be handled freely without any distinctions, and

2) Various Digital-Twin instances can be easily federated and jointly optimized, so that any Digital-Twin applications can easily utilize them to realize new services in the era of Beyond 5G.

Figure II-18.2-1 shows the proposed framework. As above discussed, target physical objects include any real-world and network objects. Other logical objects such as logical networks could also be included. Any measurement data are collected from them to reconstruct them in the Digital-Twin space, using any sensing/control devices like cameras, radio monitors, ISAC (Integrated Sensing and Communications), and legacy network managers such as syslog or EMS. Raw data analysis, such as object detection from video images, could be done here to extract meaningful information from the raw data. Then, common Digital-Twin functions would be provided by the platform so that Digital-Twin applications can utilize the Digital-Twin instances through common and open APIs. The functions include 1) device connectivity function to connect any kinds of sensing/control devices through common interfaces like WoT or MQTT, 2) device abstraction function to easily utilize various versions of devices in a common way, 3) data management functions to manage the data of any Digital-Twin instances, 4) various analysis functions, such as probabilistic inference discussed below, commonly useful for many Digital-Twin applications, and 5) Data isolation and access control function so that different applications can share the data and federate each other. Those functions may use common open source platforms like Eclipse Ditto [8].

Figure II-18.2-2 shows the implementation of Digital-Twin in distributed infrastructure. As its nature, physical devices, sensing/control functions, as well as raw data analysis are implemented on local devices or edge computing devices. Digital-Twin functions are applications are implemented on distributed computing and storage infrastructure and mutually interconnected via high-speed and low-latency communication infrastructure. We also note that all those functions shown in Figure II-18.2-1 and Figure II-18.2-2 are enabled on an advanced Beyond 5G communication, computation, and storage infrastructure.

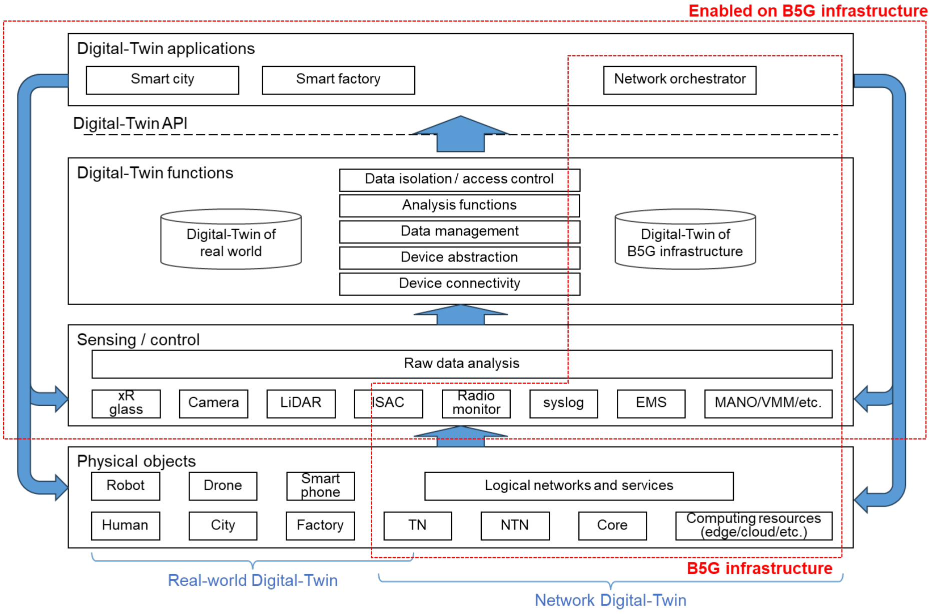


Figure II-18.2-1 Digital-Twin Platform Architecture

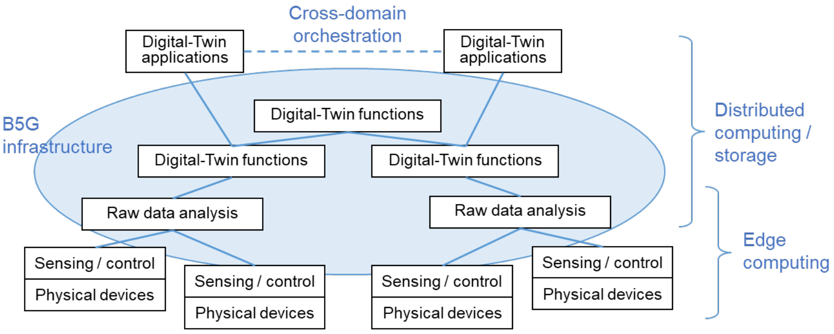


Figure II-18.2-2 Digital-Twin Implementation in Distributed Infrastructure

II-18.3. Functional Design

II-18.3.1. Probabilistic Representation of Digital-Twin

Safety and trustability are critical for many Digital-Twin applications in many real-world use cases and thus "Probabilistic Digital-Twin", in which risk management can be better handled through probabilistic representation of the real-world, and probabilistic prediction and probabilistic control based on the probabilistic representation, has been proposed [9]. Use cases of the probabilistic Digital-Twin includes following examples.

- Human Robot Collaboration, autonomous robot, automatic driving (Figure II-18.3.1-1 left): Risk sensitive path/speed/behavior control with probabilistic information for Improved/optimized/controlled safety.

- Remote-controlled robot (Figure II-18.3.1-1 center): Move slowly in unstable radio condition, avoid unstable signal area for Robust operation, avoid risks, etc.
- Network design and control (Figure II-18.3.1-1 right): BS (Base Station) location design and beam forming control based on probabilistic radio environment map for cost effective, robust, application aware network design and control.

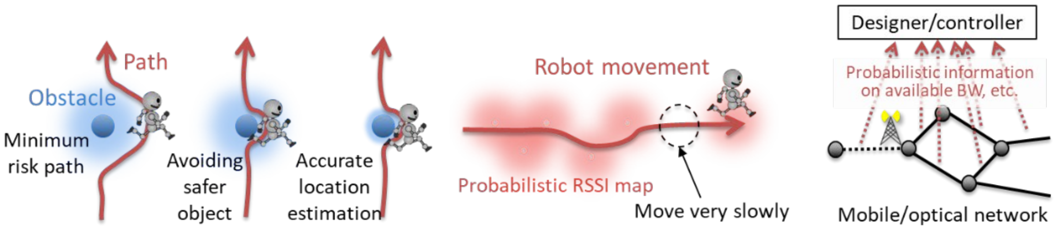


Figure II-18.3.1-1 Digital-Twin Implementation in Distributed Infrastructure

Figure II-18.3.1-2 shows an example of a data structure for probabilistic Digital-Twin. In between Digital-Twin applications and devices, a space-time structure is composed of many objects in the 4D space. These objects may be a physical object in the space (orange dot) or a status of a lattice point in the 4D space (green dot). Properties of these objects, such as occupancy status of the lattice point, location of the objects, identity/class of the objects are expressed as a probability distribution, rather than a specific value.

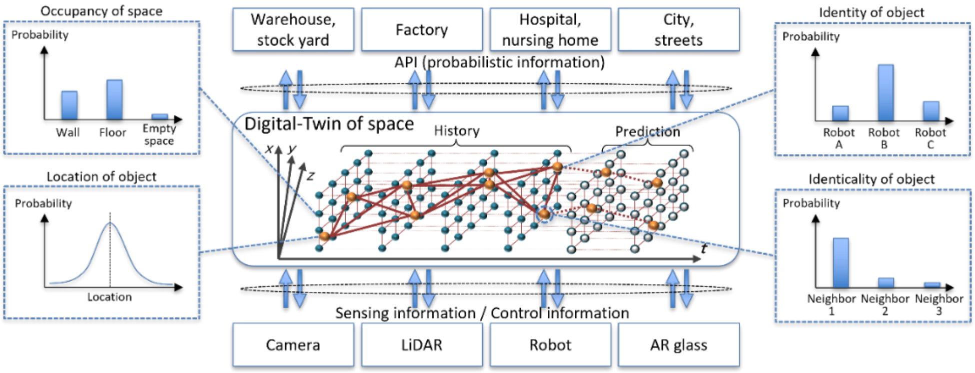


Figure II-18.3.1-2 Data Structure for Probabilistic Digital-Twin

II-18.3.2. Cross-Domain Orchestration of Digital-Twins

Today, many smart cities are introducing digital twins, using IoT sensors to collect and monitor urban data. Conventionally, they have put much effort to digitize physical space on cyber space in order to analyze and simulate the real world through data for situation monitoring and decision making. From now, we will focus on feedback to the real space to implement the results of analysis and simulation, which is called “digital-first” paradigm [10]. The digital twin collaboration will be the key to realize the digital-first services.

The beyond 5G/6G functional architecture[14] is an open platform to receive a diverse set of functions. One of the key components is the **orchestrator**, which is responsible for finding the right combination of system components and linking them together to meet the requirements from a CPS service. The orchestrator facilitates the coordination of digital twins across industries to realize a myriad of new value-added services.

To facilitate information sharing and interaction between digital twins among different domains, orchestrators are required to have the functions shown in Figure II-18.3.2-1. The **federation function** configures and manages federations of digital twins that update a shared virtual model while maintaining privacy data generated by physical objects within individual digital twins. The **translation function** facilitates the formal and semantic transformation of communications between digital twins in different domains. The **brokering function** identifies and authenticates digital twins, relays data transmission and reception, performs data filtering, real-time delivery, and guarantees delivery. The **synchronization function** synchronizes many-to-many interactions between physical space entities and cyberspace models between digital twins. **The registry function** registers and discovers digital twin components based on their feature information. International standardization of these functions is also underway [18].

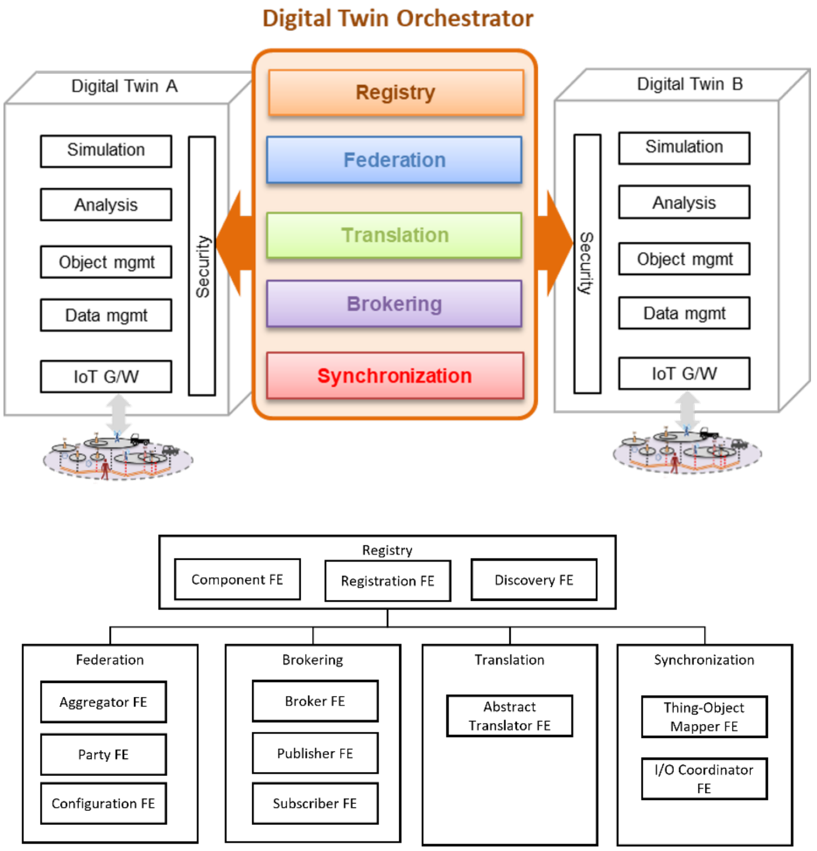


Figure II-18.3.2-1 Functional Architecture of the Digital Twin Orchestrator

II-18.4. Use Case Examples

II-18.4.1. Digital-Twin for Radio Communication Environment

This use case is to manage the radio communication network via Digital-Twin. Stability of radio communication is crucial for mission critical communications such as remote robot operations or connected car operations. However, radio communication is heavily affected by radio environments, so it is very important to understand the radio environment in detail as a Digital-Twin to manage radio communications.

As discussed in Section 3.1, it is quite difficult to monitor, estimate, and predict the radio environment, especially when high frequency radio like mm-Wave is used for mobile communications. For example, RSRP (Reference Signal Received Power) varies greatly depending on the position and angle of the terminal, as shown in Figure II-18.4.1-1, thus it should be quite useful to construct the Digital-Twin of radio environment using probability distributions, as shown in Figure II-18.4.1-2. To construct probabilistic representation of radio environment as a Digital-Twin, various statistical methods have been proposed, such as a method using Markov Random Field (MRF) [15] and Gaussian Process Regression (GPR) [16].

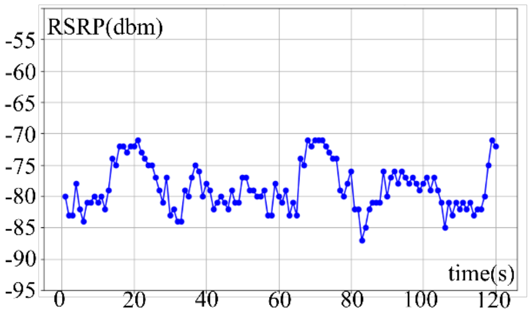


Figure II-18.4.1-1 Changes in RSRP (28 GHz Local 5G environment (Band n257), one rotation of the terminal in 50 seconds)

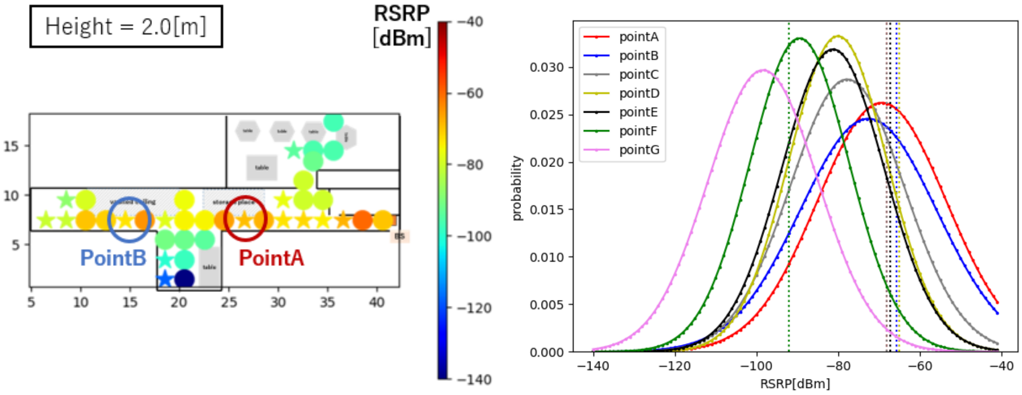


Figure II-18.4.1-2 Estimated Radio Environment Map and its Probability Distribution

Probabilistic Digital-Twin of radio environment can be used for radio network design in which estimated RSRP value should be larger than required value plus certain margin at each location. We proposed to set the margin based on the inferred probability distribution, rather than to set a uniform margin. As shown in Figure II-18.4.1-3, when the target coverage rate, i.e. the ratio of points that the observed RSRP value is within the margin, is set to 90%, the proposed method can achieve this by using a 1.1σ interval as the margin at each point, whereas the conventional method requires a uniform 1.9σ interval average as the margin. The probabilistic Digital-Twin can also be used for dynamic beam forming. As shown in Figure II-18.4.1-4, RSRP map would be estimated after beam change to see if the change is effective for the robot locations in the field, or the map would be estimated before beam change to select the best beam to satisfy the communication requirements of the robots in the field.

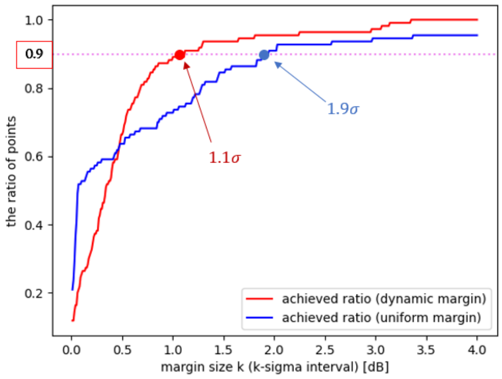


Figure II-18.4.1-3 Probabilistic Digital-Twin for radio network design (ratio of points that the observed value is within the margin)

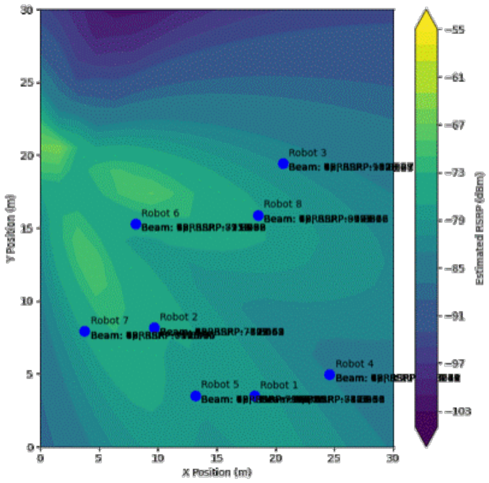


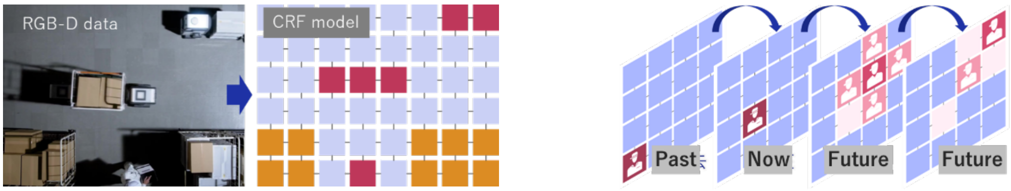
Figure II-18.4.1-4 Probabilistic Digital-Twin for dynamic beam forming (estimated RSRP map and robot locations)

II-18.4.2. Digital-Twin for Human-Robot Cooperation

Robots are widely utilized in industrial sites due to the decrease in the number of workers. However, there are many sites where it is difficult to replace all operations with robots due to cost and environmental adaptability, and robots and workers need to coexist. A logistics facility is a such site.

Logistics facilities have become larger and larger in recent years, and robots (AGVs and AMRs) are increasingly used to transfer goods inside them. On the other hand, there are many tasks that are difficult for robots to handle goods directly, for example picking and repacking goods, so manual labor is also indispensable. Therefore, workers and transfer robots coexist. Although there are some sites that separate the space for both workers and robots, it is desirable for both to be able to coexist safely in the same space to increase the efficiency of space utilization in the facility. In such cases, the trade-off between safety and efficiency becomes an issue. A typical transfer robot restrains its speed so that it can stop when an obstacle including workers approaches, and once it recognizes the obstacle, it stops. While this ensures safety, it inevitably reduces transfer efficiency. This trade-off can be resolved by utilizing probabilistic Digital Twin to predict the future location of the worker and control robots to consider the risk of collision and speed reduction. Each of the location prediction and control techniques is introduced in detail below.

However, sensors inevitably have blind spots, and there is a delay between detection of location of an obstacle and robot control, so obstacle location information at the time when the robot is operating is needed. To solve this problem, the presence or absence of obstacles at each time and point in the robot operation area is expressed as a probability, and based on the observed information, the condition of the blind spots and the future condition at each point are estimated as probabilities (See Figure II-18.4.2-1). When estimating the probability, it is important to understand the relationship of obstacles in space and time. In other words, for moving obstacles, if the obstacle is within a certain distance in the direction of movement from the point where it was observed at the previous time, the probability of its presence is high, but if it is further away than a certain distance, the probability of its presence is low. We represent such a spatial-temporal relationship between the presence and absence of obstacles as a conditional random field, CRF, and by mapping the observed values, we construct a model that predicts the future situation of obstacles from the current obstacle situation based on the observed values[11].



a: Probability of obstacle existence at each point (yellow is high, red is middle, blue is low)

b: Prediction of worker’s future location

Figure II-18.4.2-1 Prediction of future location of obstacle

After the location of future obstacles is estimated, it is necessary to control the route and speed of the transfer robot to travel safely and efficiently considering real-world uncertainties. Risk-sensitive stochastic control [12][13] solves these problems. In risk-sensitive stochastic control, the robot's motion equation is defined as a stochastic differential equation (see Figure II-18.4.2-2-a) because it represents the uncertainties that affect the robot's motion, such as hardware degradation and ground conditions, as a model. An evaluation function is used to choose optimal control inputs, and we design it to evaluate both safety and efficiency, as well as to be sensitive to risk (see Figure II-18.4.2-2-b). Although the value of the evaluation function will be a probability distribution because stochastic differential equation is used as equation of motion, it is possible to select the control that reduces both the value that the smaller is better and the variance as the optimal one. To determine the actual control inputs, various control inputs are prepared in advance, and the path and speed determined by solving stochastic differential equations are evaluated with the risk-sensitive evaluation function to select the optimal control (See Figure II-18.4.2-2-c).

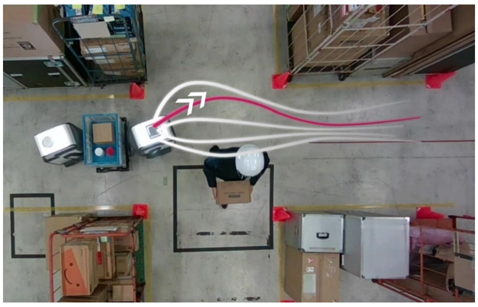
$$\underbrace{\mathrm{d}x}_{\text{State of robot}} = \underbrace{f(x,u)\mathrm{d}t}_{\text{Usual motion of equation}} + \underbrace{\sigma(x,u)\mathrm{d}B_t}_{\text{Term pf probabilistic uncertainty}}$$

a: Stochastic differential equation

$$J = \frac{1}{\beta}\log\{\mathrm{E}\left[\exp(\beta S)\right]\}$$

S : Original evaluation function
β : Risk parameter

b: Risk-sensitive evaluation function



c: Safe and efficient path

Figure II-18.4.2-2 Risk-sensitive stochastic control

II-18.4.3. Smart Sustainable Mobility

Today, the environment and mobility are major issues for many smart cities. Here we assume the following digital twin; the smart environment digital twin monitors air pollution by collecting air quality data from observation stations, while restricting emissions at major sources when air pollution is expected to worsen; the smart driving digital twin monitors the driving environment of individual cars using in-vehicle sensors, while guiding driving maneuvers and travel routes according to changes of the environment. The eco-driving assistance, an application of digital twin orchestration owned by a city officer, aims to improve the city's environmental quality by recommending environment-friendly driving maneuvers to drivers and autonomous cars in areas with poor environmental quality. Based on the emission restriction plan simulated by the smart environment digital twin, the smart driving digital twin instructs the navigation system to perform driving maneuvers to control emissions. Furthermore, it enhances the air pollution prediction of the smart environment digital twin using environmental sensor data captured by the cars, which enables more effective eco-driving assistance.

Figure II-18.4.3-1 show interactions between these digital twins through the orchestrator functions. The federation function shares the air pollution prediction model of the smart environment digital twin with the smart driving digital twin for federated learning using private data collected by individual car. The brokering function allows application to receive the emission restriction plan generated by the smart environmental digital twin, determines the restriction order for cars driving in the restricted area, and can send the order to the smart driving digital twins of the target cars. The translation function converts the environmental sensor data collected by the smart driving digital twin of individual cars to the format of observation data in the smart environment digital twin to import the “mobile” observation data for denser prediction of air pollution.

Implementation of the orchestrator framework is promoted for individual digital twin platforms as a common interface of inter-platform digital twin orchestration. The first implementation of the orchestrator framework and the use case is being conducted on NICT xData Platform [17] and Testbed. The framework implementation for IOWN is also being discussed in IOWN Global Forum based on mapping the orchestrator functions to the IOWN Data Space for Digital Twin Applications architecture [19]. In addition, integrated architecture of the orchestrator between physical space and cyber space is included in our future work.

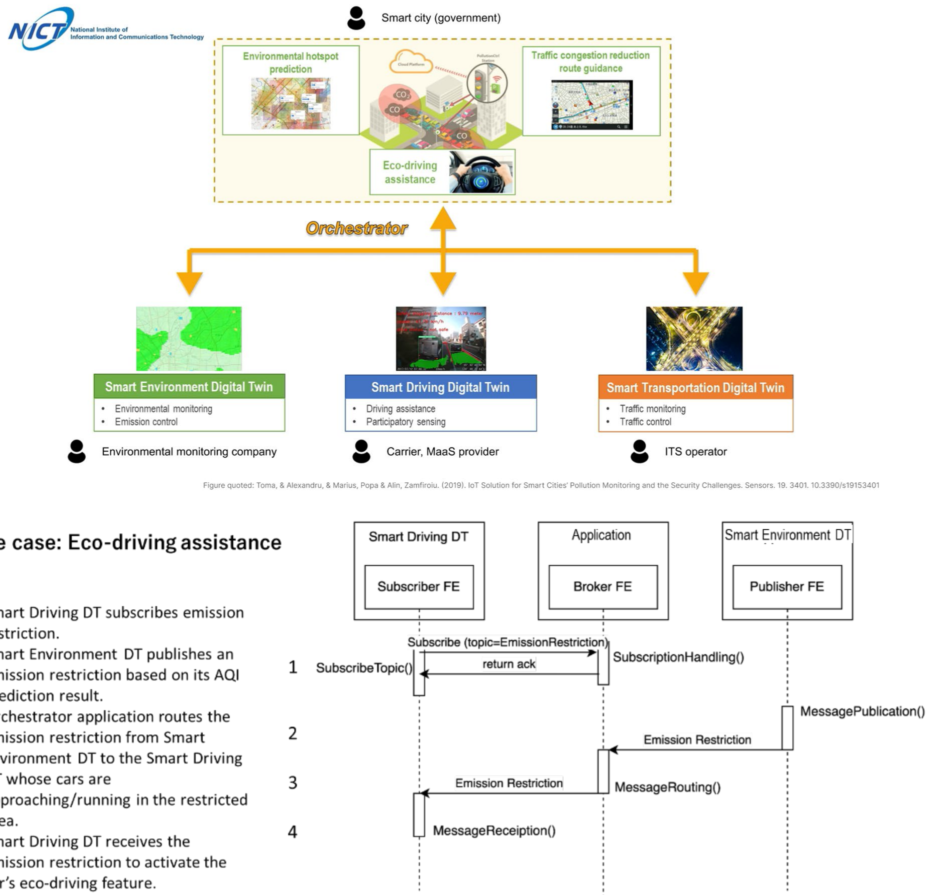

NICT National Institute of Information and Communications Technology
Smart city (government)
Environmental hotspot prediction
Cloud Platform
PollutionCtrl Station
CO_2
CO
CO
Traffic congestion reduction route guidance
Eco-driving assistance
Orchestrator
Smart Environment Digital Twin
• Environmental monitoring
• Emission control
Smart Driving Digital Twin
• Driving assistance
• Participatory sensing
Smart Transportation Digital Twin
• Traffic monitoring
• Traffic control
Environmental monitoring company
Carrier, MaaS provider
ITS operator

Figure quoted: Toma, & Alexandru, & Marius, Popa & Alin, Zamfiroiu. (2019). IoT Solution for Smart Cities' Pollution Monitoring and the Security Challenges. Sensors. 19. 3401. 10.3390/s19153401

Use case: Eco-driving assistance

1. Smart Driving DT subscribes emission restriction.
2. Smart Environment DT publishes an emission restriction based on its AQI prediction result.
3. Orchestrator application routes the emission restriction from Smart Environment DT to the Smart Driving DT whose cars are approaching/running in the restricted area.
4. Smart Driving DT receives the emission restriction to activate the car's eco-driving feature.

Figure II-18.4.3-1: Smart sustainable mobility use case

II-18.5. Conclusion

In this article, we argued that a Digital-Twin can be digital representation of both real-world network objects. Based on this, we proposed a Digital-Twin architecture which manages various Digital-Twin instances in a common way so that any Digital-Twin applications can easily utilize them. We then introduced probabilistic Digital-Twin and cross-domain orchestration of Digital-Twins, as well as the use cases including radio communication environment, human-robot cooperation, and smart sustainable mobility.

Acknowledgements

This work was partly supported by MIC under a grant entitled R&D of ICT Priority Technology (JPMI00316). Also, this research results were partly obtained from the commissioned research (No.00701) by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCE

[1] Grieves, Michael and Vickers, John. “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems”. 10.1007/978-3-319-38756-7_4., 2017.

[2] A. Fuller, et. al., "Digital Twin: Enabling Technologies, Challenges and Open Research," in IEEE Access, vol. 8, pp. 108952-108971, 2020.

[3] Y. Wu, K. Zhang and Y. Zhang, "Digital Twin Networks: A Survey," in IEEE Internet of Things Journal, vol. 8, no. 18, pp. 13789-13804, 2021.

[4] Xiaochen Zheng, et. al., “The emergence of cognitive digital twin: vision, challenges and opportunities”, International Journal of Production Research, 60:24, 2022.

[5] “Digital twin network – Requirements and architecture”, Recommendation ITU-T Y.3090, 2022.

[6] “Innovative Optical and Wireless Network Global Forum Vision 2030 and Technical Directions”, IOWN Global Forum, available at https://iowngf.org/wp-content/uploads/2023/03/IOWN_GF_WP_Vision_2030_2.0-2.pdf (Accessed: Jan. 2024).

[7] TMforum. available at https://www.tmforum.org/ (Accessed: Jan. 2024).

[8] Eclipse Ditto: Digital Twin framework of Eclipse IoT, available at https://eclipse.dev/ditto/index.html (Accessed: Jan. 2024).

[9] H. Shimonishi , D. Kominami , Y. Ohsita , H. Yoshida, K. Nogami, D. Kanetomo, and M. Murata, "Probabilistic Representation and Its Application of Digital-Twin of Spatio-Temporal Real-World Toward Trustable Cyber-Physical Interactions," in *IEEE Network*, vol. 38, no. 6, pp. 130-137, Nov. 2024.

[10] Toru Yamada "The Digital Twin and its Evolution from the International Standardization of Smart Cities.", A1-07, Interop Tokyo 2023. (2023)

[11] Y. Ohsita, S. Yasuda, T. Kumagai, H. Yoshida, D. Kanetomo, and M, Murata, “Spatio-temporal model that aggregates information from sensors to estimate and predict states of obstacles for control of moving robots,” IEICE Proceedings Series, 2022

[12] S. Yasuda, T. Kumagai and H. Yoshida, ”Cooperative Transportation Robot System Using Risk-Sensitive Stochastic Control,” 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp.5981-5988.

[13] S. Yasuda, T. Kumagai and H. Yoshida, ”Safe and Efficient Dynamic Window Approach for Differential Mobile Robots With Stochastic Dynamics Using Deterministic Sampling,” in IEEE Robotics and Automation Letters, vol. 8, no. 5, pp. 2614-2621, May 2023.

[14] Beyond 5G/6G white paper version 3.0, ISBN 978-4-904020-32-6, 2023, available at https://beyond5g.nict.go.jp/download/index.html (Accessed: Jan, 2024)

[15] D. Kodama, K. Ohira, H. Shimonishi, T. Nakahira, D. Murayama, and T. Ogawa, “Enhancing Indoor Millimeter Radio Communication: A Probabilistic Approach to RSS Map Estimation”, in Proc. of IEEE Consumer Communications & Networking Conference (CCNC), 2024.

[16] T. Bushi, T. Nakahira, S. Shinohara, Y. Asai, K. Ohira and H. Shimonishi, "Real Time Reconstruction of Radio Environment Maps in Indoor Millimeter-Wave Beamforming with Beam Changes," 2024 20th International Conference on Network and Service Management (CNSM), Prague, Czech Republic, 2024, pp. 1-7

[17] NICT xData Platform, https://www.xdata.nict.jp/en/ (Accessed: Jan, 2024).

[18] ITU Focus Group on metaverse, “High-level interoperability architecture for cross-platform metaverse”, ITU Focus Group Technical Specification, FGMV-43 June 2024.

[19] Data Space for Digital Twin Applications - Functional Architecture, IOWN Global Forum, December 2024.

II-19. Task-Oriented 6G Native-AI Network Architecture

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

***Abstract*— The vision for 6G networks is to offer pervasive intelligence and internet of intelligence, in which the networks natively support artificial intelligence (AI), empower smart applications and scenarios in various fields, and create a "ubiquitous-intelligence" world. In this vision, the traditional session-oriented architecture cannot achieve flexible per-user customization, ultimate performance, security and reliability required by future AI services. In addition, users' requirements for personalized AI services may become a key feature in the near future. By integrating AI in the network, the network AI has more advantages than cloud/MEC AI, such as better QoS assurance, lower latency, less transmission and computing overhead, and stronger security and privacy. Therefore, this article proposes the task-oriented native-AI network architecture (TONA), to natively support the network AI. By introducing task control and quality of AI services (QoAIS) assurance mechanisms at the control layer of 6G [1].**

II-19.1. Introduction

This explains the needs of Native-AI based 6G Wireless Network Architecture and lists of reason that requires to shift to task-oriented system mechanism. The proposed NW architecture called Task-Oriented native-AI network architecture (TONA), to natively support the network AI that create a "ubiquitous-intelligence" world. Reflecting the proceeding transformation, this article further proposes TONA to meet personalized AI service demand and requirements. This article mainly:

(1) Introduces three-layer logical architecture of task management and control system, and designs the task lifecycle management procedures, which include the collaboration of multi-dimension heterogeneous resources (communication, computing, data, and algorithm) and multi-node at the control layer.

(2) Defines task-specific QoAIS indicators for the mapping from Service Level Agreement (SLA) indicators — e.g., service requirement zone (SRZ) and user satisfaction ratio (USR) — to QoAIS indicators, and discusses task-level QoS assurance to meet individual requirements of different users.

(3) Compares the network AI and cloud/mobile edge computing (MEC) in terms of QoAIS indicators. Thanks to providing the AI executing environments closer to UE, TONA is anticipated to have some advantages, such as better data privacy protection, lower latency, and lower energy consumption.

II-19.2. Network Paradigm Change

The TONA, as shown in Figure II-19.2-1, introduces the orchestration and control functions as well as the resource layer in network AI. The control function uses control layer signaling to control multi-nodes (UEs, base stations, and CN NEs) and heterogeneous resources in real-time. We believe that the 6G network architecture requires the following changes in the design paradigm:

(1) Change 1: The object to be managed and controlled in network are changed from sessions to tasks.

(2) Change 2: The resources of the object are changed from one dimension to multi-dimensions, from homogeneous to heterogeneous.

(3) Change 3: The object control mechanism are changed from session-control to task-control.

(4) Change 4: The performance indicators of the object are changed from session-QoS to task-QoS.

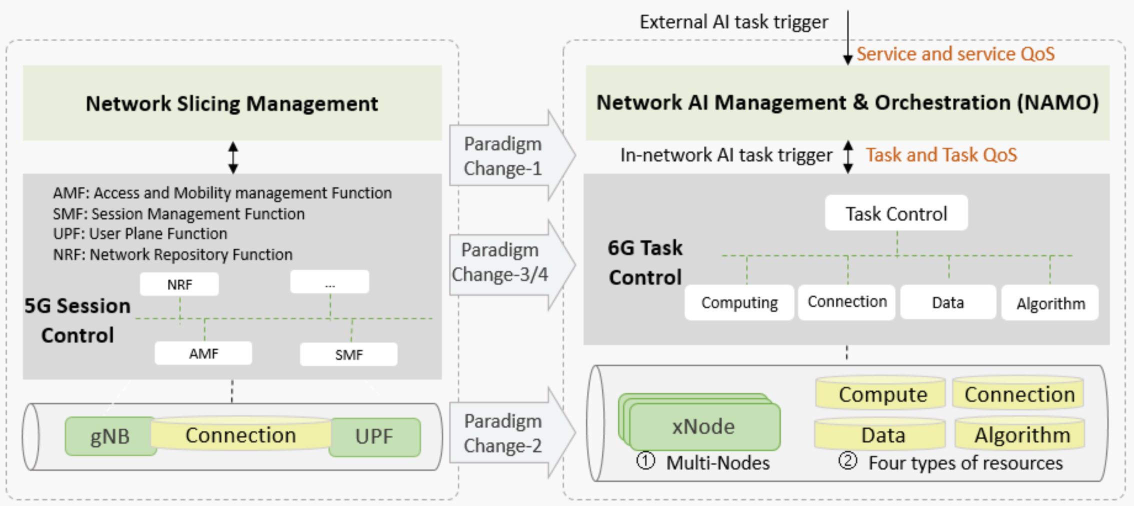


Fig. II-19.2-1 Network paradigm changes

II-19.2.1. Change 1: From Session to Taks

AI tasks differ from traditional sessions in terms of technical objectives and methods.

In terms of technical purposes, a traditional communications system provides session services, typically between terminals or between terminals and application servers, to transmit user data (including voice). Conversely, network AI (i.e., NE intelligence and network intelligence) aims to provide intelligent services for networks and improve communication network efficiency. Service intelligence seeks to provide app-specific intelligent services for third parties. Thus, sessions and AI tasks have different purposes.

II-19.2.2. Change 2: From Single-dimension to Multi-dimension Heterogenous Resources

The traditional wireless system establishes tunnels and allocates radio resources for data transmission. Conversely, TONA implements collaboration among heterogeneous resources of connection, computing, data and model/algorithm to execute AI tasks. Take an AI inference task as an example.

II-19.2.3. Change 3: From Session-control to Task-control

Unlike session control, task management and control in network AI includes the following functions: (1) Decomposing and mapping from external services to internal tasks, (2) Decomposing and mapping from service QoS to task QoS, and (3) Providing heterogeneous and multi-node collaboration mechanisms to orchestrate and control heterogeneous resources of multiple nodes at the infrastructure layer in real-time (to implement distributed serial or parallel processing of tasks and real-time QoS assurance).

II-19.3. Architecture and Key Technologies

This section describes the logical architecture and deployment options of TONA, and QoAIS details.

II-19.3.1. Logical Architecture of TONA

First, we introduce fundamental basic concepts in wireless network. A communications system consists of a management domain and a control domain. The Operations Administration and Maintenance (OAM) deployed in management domain is used to operate and manage NEs through non-real-time (usually within minutes) management plane signaling. The control domain is deployed on core network (CN) NEs, base stations, and terminals, and features with real-time controlling signaling (usually within milliseconds). For example, an E2E tunnel for a voice call can be established within dozens of milliseconds by control signaling.

Unlike the centralized, homogeneous, and stable AI environment provided by the cloud, the network AI faces the following technical challenges when embedded in the wireless networks: (1) AI needs to be distributed on numerous CN NEs, base stations, and UEs. Therefore, it is necessary to consider how to manage the massive number of nodes efficiently in the architecture design. (2) The computing, memory, data, and algorithm capabilities of different nodes vary significantly, requiring the architecture design to also consider how to efficiently manage these heterogeneous nodes efficiently. (3) The dynamic variation of the channel status and the computing load need to be factored into the architecture design. To address the aforementioned challenges, TONA includes two logical functions, as shown in Figure II-19.3.1-1: (1) AI orchestration and management, called Network AI Management & Orchestration (NAMO); and (2) task control. NAMO decomposes and maps AI services to tasks and orchestrates the AI service flows. It is not performed in real-time and is generally deployed in the management domain. Task control introduces the Task Anchor (TA), Task Scheduler (TS), and Task Executor (TE) functions in the control domain in three layers. This layered design strikes a balance between the task scope and real-time task scheduling, and effectively manages the numerous, heterogeneous nodes and aware of dynamic change of heterogeneous resources (e.g. channel status and computing load).

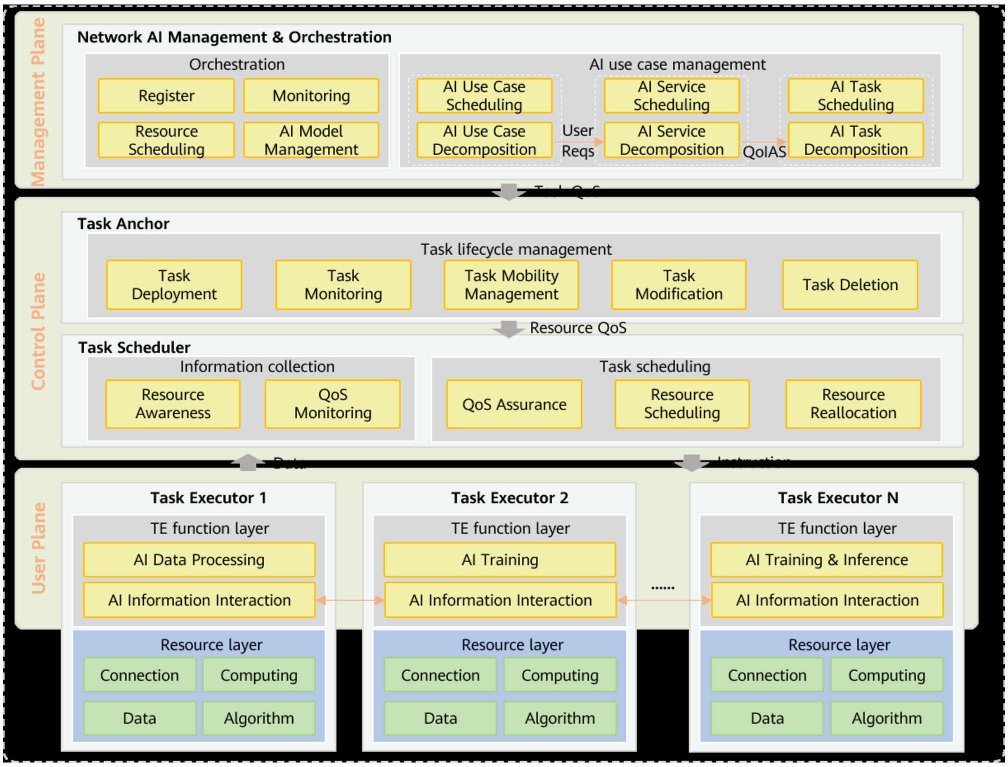


Fig. II-19.3.1-1 Logical architecture of TONA

II-19.3.2. Deployment Architectures

The statuses of TEs (e.g., the CPU load, memory, electricity, and UE channel status) change in real-time. As such, deploying TA and TS close to each other can reduce the management delay. According to the design logic of wireless networks, the CN and RAN need to be decoupled as much as possible. For example, the CN should be independent of RAN Radio Resource Management (RRM) and Radio Transmission Technology (RTT) algorithms. Therefore, this article recommends that TA/TS be deployed on RAN and CN, named RAN TA/TS and CN TA/TS, respectively. This way will allow TA to manage TEs in real-time flexibly. Four deployment scenarios of TONA are shown in Figure II-19.3.2-1 to describe the necessity and rationality of CN TA and RAN TA. These scenarios are only examples — there may be other deployment scenarios and architectures.

Scenario 1: gNodeB + UEs. In this scenario, the gNodeB serves as both TA and TS, and the UEs serve as TEs. Here, a UE is a computing provider and task executor, which accepts task assignment and scheduling from the gNodeB. The Uu interface and Radio Resource Control (RRC) layer between the gNodeB and the UE can be enhanced to support task controlling and scheduling purposes.

Scenario 2: CU + DUs. In this scenario, the CU serves as both TA and TS, and the DUs serve as TEs. Here, a DU is the computing provider and task executor. The F1 interface and F1-AP layer between the CU and the DU can be enhanced to support task controlling and scheduling purposes.

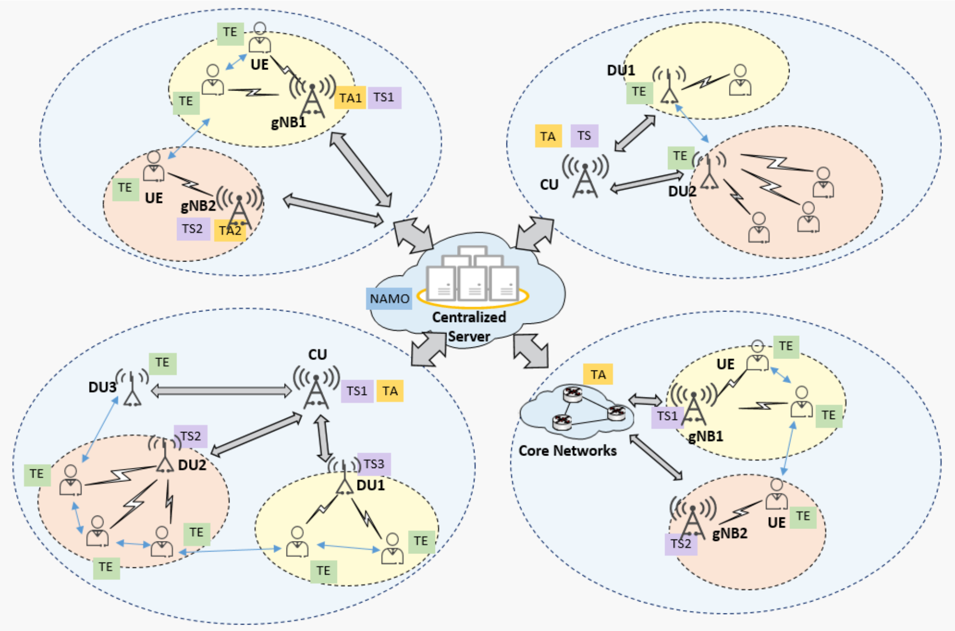


Fig. II-19.3.2-1 Four deployment scenarios of TONA

II-19.4. Advantage Analysis

Compared with cloud/MEC AI, the TONA and QoAIS have the following advantages (summarized in Table 2) in meeting users' customized AI service requirements:

(1) QoAIS assurance

Dynamic wireless environments require joint optimization of the heterogeneous resources (connection and three AI resources) to achieve precise QoAIS assurance.

(2) Latency

TONA computing is distributed on NEs closer to UEs or even directly on UEs to process data locally. This not only successfully achieves real-time and low-latency AI services, but also significantly reduces data transmission. In the cloud/MEC AI mode, a large amount of data needs to be transmitted to the cloud/MEC for training, meaning that E2E data transmission takes longer to complete.

(3) Overhead

TONA can optimally allocate resources through the real-time collaboration mechanism of the heterogeneous resources, maximizing the overall resource utilization and reducing the transmission and computing overheads. Conversely, because the cloud/MEC AI cannot adapt to dynamic environments, it allocates resources based on only the maximum resource consumption to ensure QoAIS. As a result, the overall resource utilization is low, and the resource overhead is high.

(4) Security

TONA has native data security and privacy protection capabilities because it processes data inside the network. Unlike TONA, the cloud/MEC AI protects data privacy only at the application layer.

II-19.5. Conclusion

To meet the 6G vision of pervasive intelligence and internet of intelligence, TONA is proposed to support efficient collaboration of heterogeneous resources and multi-node in wireless networks, and to provide new services in the form of tasks at the network layer. By bringing new dimensions of resources to 6G networks (i.e., computing, data, and model/algorithm), this architecture enables the SLA assurance of computing related services such as AI services, further explores the application scenarios of 6G networks, and enriches the value of wireless networks. Furthermore, the task concept and TONA proposed in this article support not only AI tasks, but also sensing-, computing- and data processing-specific tasks.

REFERENCE

[1] IEEE Network, “Task-Oriented 6G Native-AI Network Architecture”, October 2023,
https://ieeexplore.ieee.org/document/10273257

Abbreviation List

Abbreviation	Explanation
3GPP	3rd Generation Partnership Project
5G	5th Generation mobile communication systems
6G	6th Generation mobile communication systems
AA	Actor Allocator
Adam	Adaptive Moment Estimation
AGV	Automatic Guided Vehicle
AI	Artificial Intellegence
AI-AI	Ai-native Air Interface
AM	Amplitude Modulation
AMR	Autonomous Mobile Robot
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
BER	Bit Error Rate
BF	Beamforming
BLER	Block Error Rate
BM	Beam Management
bMRO	beam-based Mobility Robustness Optimization
BS	Base Station
BSS	Business Support System
CDF	Cumulative Distribution Function
CF-mMIMO	Cell-free massive MIMO
CIR	Channel Impulse Response
CLC	Closed-Loop Control
CMOS	Complementary Metal-Oxide-Semiconductor
CN	Core Network
CNN	Convolutional Neural Network
CPS	Cyber-Physical System
CPU	Central Processing Unit
CSI	Channel State Information

Abbreviation	Explanation
CSI-RS	Channel State Information Reference Signal
CU	Central Unit
DC	Direct Current
DCNN	Deep Convolutional Neural Network
DDQN	Double Deep Q Network
D-DRL	Distributed DRL
DL	Downlink
DM-RS	Demodulation Reference Signal
DNN	Deep Neural Network
DPD	Digital Predistortion
DRL	Deep Reinforcement Learning
DSP	Digital Signal Processing
DT	Digital Twin
DU	Distributed Unit
eBPF	extended Berkley Packet Filter
eMBB	enhanced Mobile Broadband
EMS	Electronics Manufacturing Service
ES	Energy Saving
EVM	Error Vector Magnitude
FC	Fully Connected
FDE	Frequency Domain Equalization
FFT	Fast Fourier Transformation
FL	Federated Learning
FNN	Fully connected Neural Network
FPGA	Field Programmable Gate Array
GA	Genetic Algorithm
gNB	gNodeB
GNN	Graph Neural Network
GoB	Grid of Beams
GPR	Gaussian Process Regression
GPU	Graphics Processing Unit

Abbreviation	Explanation
Grand-CAN	Gradient-weighted Class Activation Mapping
HDD	Hard Disk Drive
HLF	Horizontal Federated Learning
HSL	Horizontal Split Learning
IBO	Input Back Off
IFFT	Inverse FFT
IMT	International Mobile Telecommunication
INL	In-Network Learning
IoT	Internet of Things
IOWN	Innovative Optical and Wireless Network
ISAC	Integrated Sensing And Communications
ITU-R	International Telecommunication Union-Radiocommunication Sector
KPI	Key Performance Indicator
LAN	Local Area Network
LCM	Life Cycle Management
LiDAR	Light Detection And Ranging
LLM	Large Language Model
LLR	Log-Likelihood Ratio
LMF	Location Management Function
LOS	Line-Of-Sight
LQ	Link Quality
MCS	Modulation and Coding Scheme
MDP	Markov Decision Process
MDT	Minimization of Drive Tests
MEC	Mobile Edge Computing
MEC	Multi access Edge Computing
MIMO	Multiple input Multiple Output
ML	Machine Learning
MLP	Multi-Layer Perceptron
mMIMO	massive MIMO
mmWave	millimeter Wave

Abbreviation	Explanation
MP	Memory Polynomial
M-plane	Management plane
MPLS-TE	Multiprotocol Label Switching - Traffic Engineering
MQTT	Message Queuing Telemetry Transport
MRF	Markov Random Field
MS	Mobile Station
MU-MIMO	Multi-User MIMO
MVP-C	Minimum Viable Plan Committee
NAMO	Network AI Management and Orchestration
NE	Network Element
Near-RT RIC	Near-Real-Time RIC
NLOS	Non-Line-Of-Sight
NN	Neural Network
Non-RT RIC	Non-Real-Time RIC
NR	New Radio
NRNT	New Radio Network Topology
NSSMF	Network Slice Subnet Management Function
NW	Network
OAM	Operations, Administration, Maintenance
O-CU-CP	O-RAN Central Unit - Control Plane
O-CU-UP	O-RAN Central Unit - User Plane
O-DU	O-RAN Distributed Unit
OFDM	Orthogonal Frequency Division Multiplexing
OLPC	Outer-Loop Power Control
O-RAN	Open Radio Access Network
O-RU	Open RAN Radio Unit
OSS	Operation Support System
OTT	Over-The-Top
PA	Power Amplifier
PF	Proportional Fairness
PGW	Packet data network Gateway

Abbreviation	Explanation
PL	Path Loss
PM	Performance Metric
PoC	Proof-of-Concept
PS	Parameter Server
Q1	the first Quarter
QAM	Quadrature Amplitude Modulation
QoAIS	Quality of AI Services
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
R&D	Research and Development
RAN	Radio Access Network
rApp	RAN intelligent controller Application
RAT	Radio Access Technology
ResNet	Residual Network
RF	Radio Frequency
RF	Random Forest
RIC	RAN Intelligent Controller
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
RMS	Root Mean Square
RMSE	Root Mean Square Error
ROS	Robot Operating System
RRM	Radio Resource Management
RS	Relay Station
RSRP	Reference Signal Received Power
RSSI	Received Signal Strength Indicator
RT	Ray Tracing
RTT	Radio Transmission Technology
RVTDNN	Real-Valued Time-Delay Neural Network
Rx	Receiver

Abbreviation	Explanation
SA	Static Approach
SB	Subband
SC	Single Carrier
SCS	Subcarrier Spacing
SGCS	Squared Generalized Cosine Simularity
SINR	Signal-to-Interference plus Noise Ratio
SL	Split Learning
SLA	Service Level Agreement
SMO	Service Management and Orchestration
SNR	Signal-to-Noise Ratio
S-NSSAI	Single-Network Slice Selection Assistance Information
SON	Self Organizing Network
SOTA	State-Of-The-Art
SRZ	Service Requirement Zone
SSB	Synchronization Signal Block
STA	Station
SVM	Support Vector Machine
TA	Task Anchor
TAT	Turn Around Time
TBD	To Be Determined
TDD	Time Division Duplex
TDL	Tapped Delay Line
TE	Task Executor
TR	Technical Report
TRP	Transmission and Reception Point
TS	Task Scheduler
Tx	Transmitter
UCTG	Use Case Task Group
UE	User Equipment
UL	Uplink
UMa	Urban Macro

Abbreviation	Explanation
UPF	User Plane Function
UPT	User Perceived Throughput
USR	User Satisfaction Ratio
vCPU	Virtualized CPU
VFL	Vertical Federated Learning
VNF	Virtualized Network Function
VR	Virtual Reality
vRAN	virtual RAN
VSL	Vertical Split Learning
WB	Wideband
WG	Working Group
WLAN	Wireless Local Area Network
WoT	Web of Things
WP	Working Party
XAI	Explainable AI
xAPP	eXtended Application
XGMF	XG Mobile Promotion Forum