O-RAN next Generation Research Group (nGRG)

Research Report

Research Report on Native and Cross-domain AI: State of the art and future outlook

Report ID: RR-2023-03

Contributors:

Keysight

AsiaInfo

China Telecom

China Unicom

Ericsson

Lenovo

Nokia

NVIDIA

Reliance Jio

VTT

ZTE

Release date: 2023.09

Authors

Author	Company
Balaji Raghothaman	Keysight Technologies (Editor-In-Chief)
Zhanwu Li	AsiaInfo
Zexu Li	China Telecom
Tingting Liang	China Unicom
Mirko D'Angelo	Ericsson
Mingzeng Dai	Lenovo
Henning Sanneck	Nokia
Lopamudra Kundu, Xingqin Lin	NVIDIA
Ravi Sinha	Reliance Jio
Tao Chen	VTT
Jiajun Chen	ZTE

Disclaimer

Copyright

Executive summary

This document provides a broad view of the functional aspects that needs to be considered for the incorporation of native and cross domain AI into next generation networks. It begins with a brief overview of the current status of AI in global standards organizations, including 3GPP, O-RAN and ETSI-ZSM. The report provides concise definitions of the terms native and cross domain AI in the context of wireless networks and then goes on to discuss the impact of AI on the architecture.

The challenges of ingesting large amounts of disparate data across multiple layers, and the corresponding requirements on data modeling, formatting and representation are discussed. A unified data ingestion model is emerging as a key requirement. The importance of distributed and edge intelligence to solve the complex multi-layered issues in wireless networks is emphasized, along with the notion of trustworthiness in such a distributed architecture. Enablers for large scale distributed intelligence, including HPC platforms and accompanying software platforms including open-source, are discussed. The emerging paradigm of intent-driven management, and its interplay with AI/ML are described. The necessity to have collaborative AI across disaggregated RAN and between RAN and CN are discussed.

This research report is the first attempt in O-RAN nGRG to survey the landscape with respect to AI/ML as it applies to next generation networks, and provides a foundation based on which several further explorations into each of the highlighted areas can be initiated.

Table of Contents

List of abbreviations

AAL	- Abstraction and Acceleration
CNN	- Convolutional Neural Network
DNN	- Deep Neural Network
GAN	- Generative adversarial networks
IaaS	- Infrastructure as a Service
MnS	- Management Services
MD	- Management Domain
MDAS	- management and data analytics service
NWDAF	- Network Data Analytics Function
PaaS	- Platform as a Service

List of figures

1. Background

1.1 Overview of AI in wireless networks

As we move through successive generations of wireless technology, the complexity of the network has increased exponentially, making it harder to create and solve analytical models to describe the behavior of various subsystems or the system as a whole. Methodologies based on machine learning and artificial intelligence are ideally suited for such scenarios, and thus it is inevitable that we see a progressively larger role for AI/ML in the network. AI/ML has already been in use for a few years in various aspects of the network, such as analytics. As we progress into 5G-Advanced, we see its usage in the Radio Access Network (RAN), Radio Resource Management (RRM) and Core Network (CN) functions as well. The introduction of the RAN Intelligent Controller (RIC) in the O-RAN architecture has been an important development, making it possible to introduce AI/ML-based solutions to a wide variety of use cases.

1.2 Overview of AI life cycle management

As the integration of AI and networks gradually deepens, the current industry consensus is that native AI will be one of the core features of the future network. AI will be ubiquitous in the future network, which will bring many profound impacts and challenges to the future network design. The life cycle management of AI will also become an important aspect of the future network.

For 5G network intelligent applications, most of the links in the AI workflow are processed offline, separate from the network function workflow, and are only utilized as an additional component of the network function. All links are developed independently for each distinct intelligent application scenario. There is a lack of coordination and sharing of resources between different scenarios, which is inefficient and costly. For the future network, the integration of native AI and network will require unified resources, operation environment and life cycle management services for end-to-end links of AI workflow in different scenarios, and this will improve the efficiency of resource utilization and raise the level of network intelligence [1], [42] .

In general, for intelligent application the AI life cycle should consist of training phase, deployment phase, and inference phase in terms of stage division, which is defined in Third Generation Partnership Project (3GPP) TR 28.908 as shown in Figure 1.2-1 is also applicable in the future network.

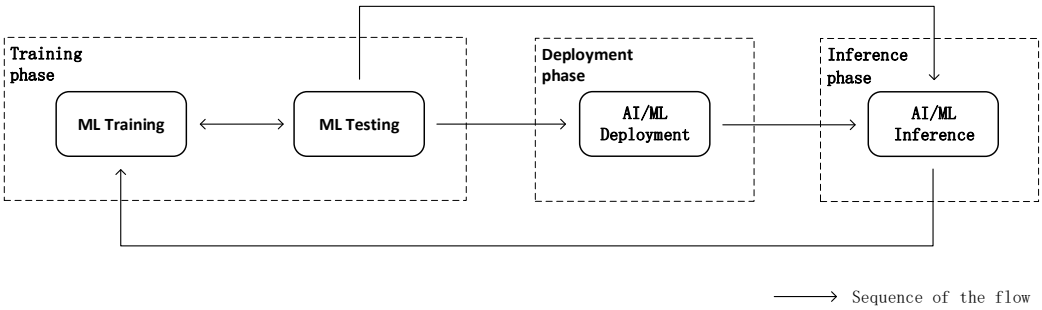

Figure 1.2-1 AI/ML Life Cycle

NOTE: Figure 1.2-1 is from 3GPP TR 28.908, Figure 4.2.1-1: AI/ML operational workflow.

From a vertical perspective, the AI life cycle management should consist of Management in ML Training phase, Management in AI/ML Deployment phase, and Management in AI/ML Inference phase in terms of stage division:

- **Management in ML Training phase:** the management function in ML Training phase may include Training Data related management, ML Training related management, ML Model Testing related management, ML Model data related Management etc.
- **Management in AI/ML Inference phase:** the management functions in AI/ML Inference phase may include Inference Data related management, AI/ML Inference task related management, Analysis Data related management etc.

From another perspective, basically, all the AI/ML applications' AI life cycle consists of three basic elements: data, AI models, and computing resources. In this way, the AI life cycle management of future network may be divided into AI Data Management, AI Model Management, AI/ML Computing Resources Management, AI/ML Inference Management, and AI/ML Service Management from a horizontal perspective, as shown in the following figure:

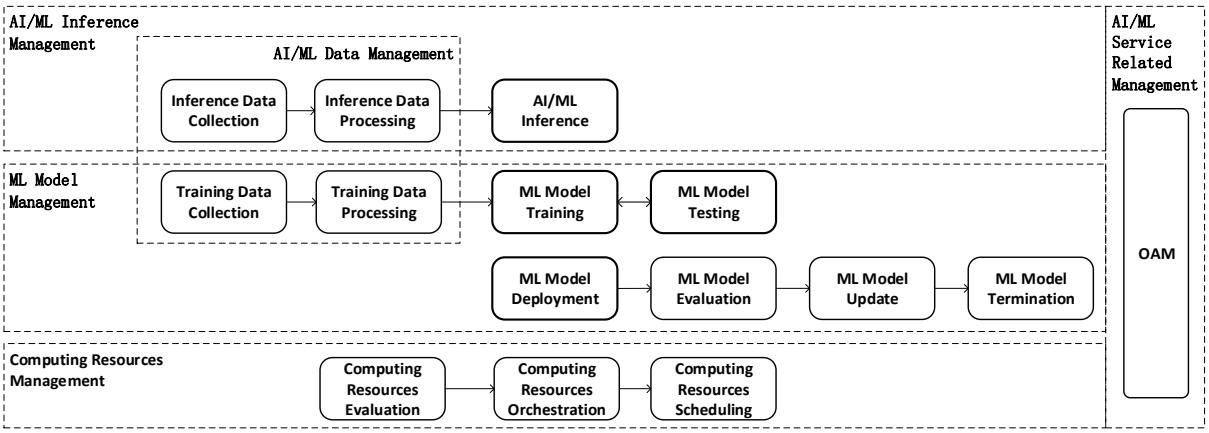

Figure 1.2-2 Horizontal perspective of AI/ML life cycle management

- **AI/ML Data Management:** The AI/ML Data mainly may include Training Data, Inference Data, and Analysis result Data etc and the AI/ML Data Management procedure may include Data Collection, Data Processing, Data Storage, Data Exposure, and Data Cleaning etc.
- **ML Model Management:** The ML Model Management procedure may include ML Model Testing, ML Model Deployment, ML Model Evaluation, ML Model Update, and ML Model Termination etc.
- **AI/ML Computing Resources Management:** The AI/ML Computing Resources Management may include AI/ML Computing Resources Evaluation, AI/ML Computing Resources Orchestration, and AI/ML Computing Resources Scheduling etc.
- **AI/ML Inference Management:** The AI/ML Inference Management may include AI/ML Inference characteristic based on the distinct intelligent application scenario.
- **AI/ML Service-Related Management:** The AI/ML Service-Related Management may include general OAM (e.g. FCAPS etc.).

1.3 Status of architecture and standardization

3GPP

3GPP has been actively working on the development of technical reports and specifications for the integration of artificial intelligence (AI)/machine learning (ML) into the fifth generation (5G) mobile networks. The work on AI/ML within 3GPP is aimed at exploring the potential use cases and technical requirements for integrating AI/ML into different parts of the 5G mobile network, including RAN, CN and services. This includes investigating how AI/ML techniques can be used to optimize network performance, improve energy efficiency, and enhance user experience.

For the radio access network, 3GPP carried out a study on further enhanced data collection for RAN in its Release 17 phase. The study identified a set of high-level principles, introduced a functional framework for RAN intelligence, and investigated the benefits of AI/ML-enabled NG-RAN with a focus on three use cases: network energy saving, load balancing, and mobility optimization. The outcome of the study can be found in the technical report TR 37.817 [3]. After finishing the study in Release 17, 3GPP is now working on a work item about AI/ML for the next-generation RAN (NG-RAN) (aka. 5G RAN). This work item is expected to introduce enhancements of data collection and signaling to support AI/ML-based network energy saving, load balancing, and mobility optimization.

Besides the normative work on AI/ML for NG-RAN in Release 18, 3GPP is carrying out a study on AI/ML for 5G New Radio (NR) air interface [4]. The study aims to develop a 3GPP framework using AI/ML for air interface, with a focus on three use cases: CSI feedback, beam management, and positioning. The study scope includes

the characterization of the defining stages of AI/ML related algorithms and associated complexity, identification of various levels of collaboration between user equipment (UE) and 5G gNB for the selected use cases, datasets for training, validation, testing and inference, and life cycle management of AI/ML models (e.g., model training, model deployment, model inference, model monitoring, model update), and common notation and terminology for AI/ML related functions, procedures and interfaces. The study involves extensive performance evaluation to assess the performance benefits of AI/ML based algorithms for air interface. The study will also assess potential specification impact to pave the way for the development of technical specifications for using AI/ML to enhance performance and/or reduce complexity or signaling overhead in air interface.

For the core network, 3GPP has been working on the development of the Network Data Analytics Function (NWDAF) as a key component of the 5G network architecture since Release 15 [5]. The NWDAF is responsible for collecting and analyzing network data to provide insights into network performance, resource utilization, and user behavior. This information can be used to optimize the performance of the network, improve the user experience, and enable the development of new applications and services that take advantage of the unique capabilities of the 5G network.

For the Operations, Administration, and Maintenance (OAM), 3GPP has been working on the development of the Management Data Analytics Service (MDAS) as part of the OAM framework for the 5G network since Release 15 [6]. The MDAS is a service that provides analytics of network data and network management information to support OAM operations, including network planning, monitoring, optimization, and troubleshooting. The MDAS is designed to collect and analyze data from various sources in the network.

Building on the MDAS results, 3GPP SA5 has studied AI/ML Management [8]. In addition to the AI life cycle management capabilities, respectively for the training and inference phases, the following management capabilities which are common to the training and inference phases were defined: AI/ML configuration and performance management as well as AI/ML trustworthiness management. The latter focuses on allowing the management service consumer to configure, monitor and evaluate the trustworthiness of an ML entity covering the whole life cycle.

A study on Security and Privacy of AI/ML-based Services and Applications in 5G is under discussion in 3GPP SA3, that aims at identifying key issues and solutions in order to address the security aspects of employing AI/ML techniques in RAN [9]. Several key issues have been identified for study: user privacy of RAN AI/ML framework, Robustness of the RAN AI/ML framework against data poisoning attacks, and security of information transfer between RAN AI/ML framework nodes.

In summary, the goal of the AI/ML work in 3GPP is to drive innovation and improve the performance of 5G and beyond by leveraging the power of AI/ML techniques [7]. This will enable the development of new applications and services, as well as enhancing the overall efficiency and effectiveness of mobile networks.

O-RAN

Figure 1.3-1 shows the proposed/ongoing study and work items related to AI/ML in 3GPP 5G-Advanced as well as O-RAN. Basically AI/ML techniques will be enabled in all parts of the system. Besides training & inference on the devices (and collaboratively across devices and the network), the NG-RAN, the Core and the Management domains contain frameworks for data collection as well as ML training/inference. Use cases in the RAN domain are for example traffic steering, load balancing and mobility optimization. AI/ML applied to the Management domain covers e.g., energy saving, and fault management (anomaly detection/diagnosis) across NG-RAN and Core. Furthermore, the Management domain also contains “Management of AI/ML” services to manage e.g., the AI/ML life cycle, training and inference execution across NG-RAN, Core and the Management domains.

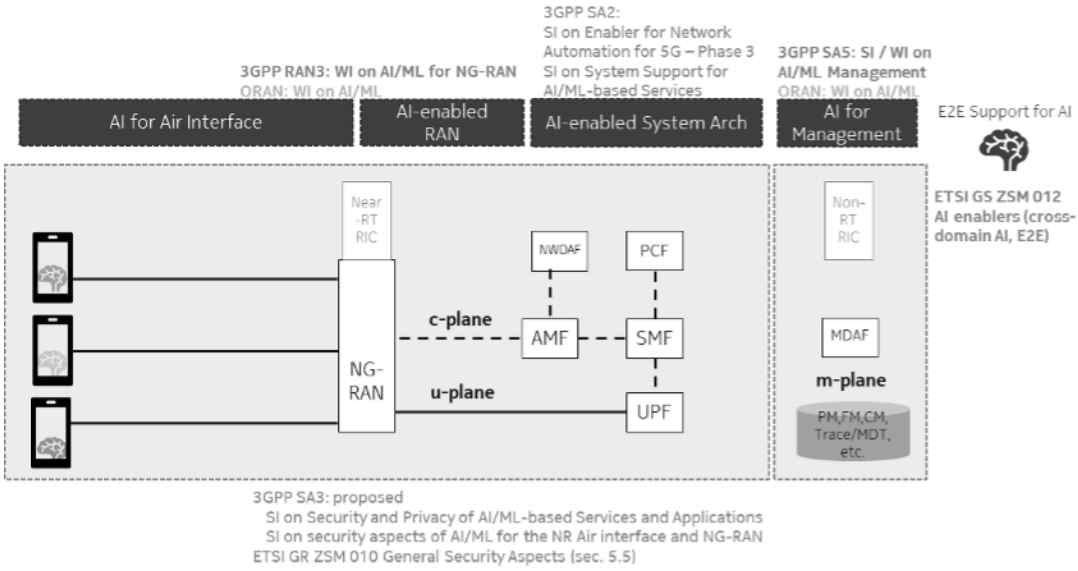

Figure 1.3-1 Overview on AI/ML-related study and work items in 3GPP and O-RAN (RAN, Core, Security, Management)

Technical report [10] introduced basics (terminology, AI/ML algorithm outline), definition of ML pipelines and their life cycle management and discussed embedding ML into the O-RAN architecture as well as ORAN ML deployment scenarios.

Recently, a work item on “AI/ML in O-RAN” [11] has been proposed and adopted which aims to analyze ML workflow related functions in the near-RT RIC architecture and specify the corresponding changes to O-RAN interfaces and APIs.

While the work to introduce a general AI/ML framework into the O-RAN architecture is still ongoing, designs have been proposed and an example design has been introduced in [10] . **Figure 1.3-2** shows basic building blocks and life cycle of AI/ML models in such a design. The design principle is to host the AI/ML model training and model repository at Service Management and Orchestration (SMO) and share them with other components in the RAN. The ready-to-use models can be deployed and utilized by rApps, xApps and network functions in the Non-RT RIC, Near-RT RIC, CU,

DU and RU. In this design, AI/ML models are designed by ML toolkits, e.g. scikit-learn, R, H2O, Keras, or TensorFlow. Initial AI/ML models are imported into SMO for training, while the training data are collected from different sources at RAN system. This design allows the unified management and sharing of AI/ML models among different near-RT RICs and RAN components.

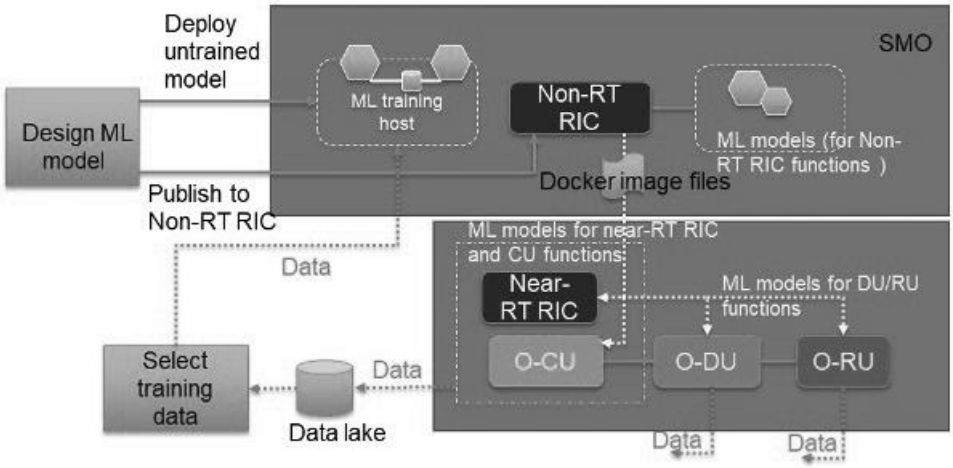

Figure 1.3-2 AI/ML framework implementation proposal in O-RAN [10]

ITU-R

ITU-R published a comprehensive report on future technology trends for terrestrial systems [12] . The ITU-R's technology report has been a key milestone in past generations of wireless specifications, including IMT-2000, IMT-Advanced, and IMT-2020, and hence this latest report acquires great importance for the next generation.

In the section devoted to technologies for AI-native communications, the aspects of AI-native air interfaces and AI-native radio network are explored. Under each aspect, some key areas of investigation are presented, which are listed below.

AI-native Air Interface

- Symbol detection/decoding
- Channel Estimation
- MAC Layer design
- Radio Resource Management
- Semantic Communications

AI-native radio network

- Intelligent data perception
- User feedback
- Pervasive computation nodes
- Supply of on-demand capacity
- Collaboration of sensing and AI
- Distributed and unified AI control

- Adaptive solutions for different usages

Radio Network to Support AI services

- Shift from DL-centric to UL-centric radio
- Shift from core network to deep edge
- Shift from cloudification to ML

Others

The goal of ETSI ZSM is to enable Zero-touch automated network and service management in a multi-vendor environment. The ZSM architecture (Figure 1.3-3) enables exposing and consuming Management Services (MnS) across a set of Management Domains (MDs) and an E2E Service Management Domain.

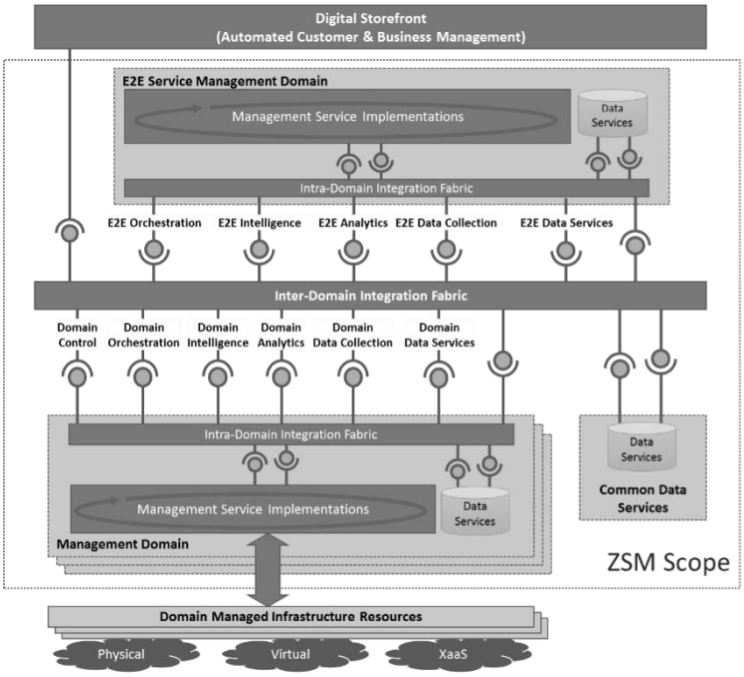

Figure 1.3-3 ETSI ZSM Architectural Framework [13]

The ZSM technical specification [14] extends the set of Management Services available within the ZSM architecture to address AI/ML aspects. In addition to common Management Services (ML event notification, log collection, feasibility check, data processing, training reporting, model cooperation management), this is done along a taxonomy of AI/ML enablers (Figure 1.3-4):

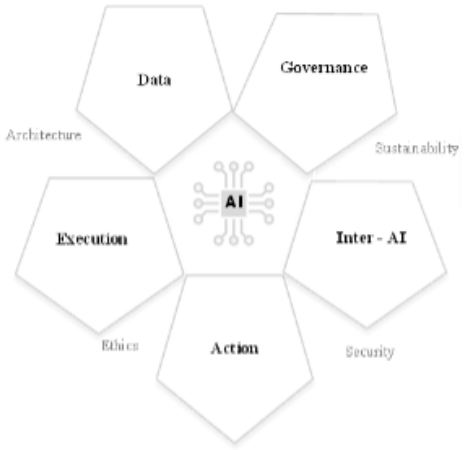

Figure 1.3-4 AI Enabling Areas [14]

- Data: providing data access across domains and ensuring e.g., data quality is crucial for AI/ML
- Execution: providing the deployment platform (compute) and operation for executing AI/ML applications (MnS: model validation, sandbox configuration)
- Action: converting the output of AI/ML applications to actions to be executed on network / management functions, management domains, etc.
- Inter-AI: coordinating interactions with the potential many types and huge number of AI/ML application instances is an important management functionality (MnS: Federated Learning (FL) configuration)
- Governance: interfacing the AI/ML-enabled management domains to the human network operator with management services targeting the mapping of intents to AI/ML applications on the one hand, and instrumenting AI/ML applications to be “trustworthy”, i.e., be explainable, robust and fair (MnS: Data & Model Trust management, Data & Model Trust evaluation, ML Fallback Management)

Thereby, ZSM is able to enable a range of “AI/ML for Network Management” as well as “Management of AI/ML” scenarios like (cf.[14], Annex A) ML-based Anomaly Detection, Federated Learning for Network Management, Trustworthy ML, Distributed ML, ML model validation and ML model cooperation.

2. Definition of Native and Cross-Domain AI

2.1 Native AI

Artificial intelligence (AI) is envisioned to be the cornerstone of 6G evolution. Continuing the transformational journey of 5G telecommunications network beyond ‘communication-only’ fabric, 6G will continue to enable further emerging verticals and new use cases, ushering advanced capabilities of pervasive intelligence throughout the network. Such omnipresence of intelligence, in turn, would imply access to AI services anytime, anywhere and by anyone [15].

Enabling AI-driven networking requires a paradigm shift in the architectural blueprint, in particular, making the radio access component of the network infrastructure built natively for AI augmentation and AI infusion.

AI-augmented RAN will make overall RAN operations and management inherently intelligent by enhancing existing NG-RAN architecture with AI/ML capable logical functions. Towards that goal, open RAN has already been marching on, with RAN Intelligent Controller (RIC) functions supported by O-RAN architecture in two flavors – non-real time and near-real time. To meet the requirements of 6G use cases with potentially more stringent throughput and/or latency requirements, existing AI/ML capabilities of O-RAN can be further augmented by RIC evolution (e.g., with the introduction of real-time RIC [16]) and RIC enhancement (e.g., by enabling intent-based networking [17]).

AI-augmented RAN can be further evolved towards AI-infused RAN, by embedding AI/ML-based submodules into existing NG-RAN nodes, potentially enhancing or even replacing traditional signal processing implementation blocks. One such example is O-DU layer 1 (L1) infused with neural network (NN) based processing blocks [18] as shown in Figure 2.1-1 below. Other potential examples include further disaggregation of existing O-RAN nodes and additional interface specifications to create flexible ‘AI-infusion’ points and enable richer data collection capability, AI-infused O-CU/O-DU control planes, accelerated data retrieval from network nodes, etc.

In a nutshell, native AI centric network design will drive evolution of O-RAN architecture towards further augmentation of RAN intelligence and enablement of AI infusion capability into logical nodes and interfaces of network architecture.

Native AI

Embedding artificial intelligence into functionalities supported by various nodes/endpoints and interfaces within a network architecture

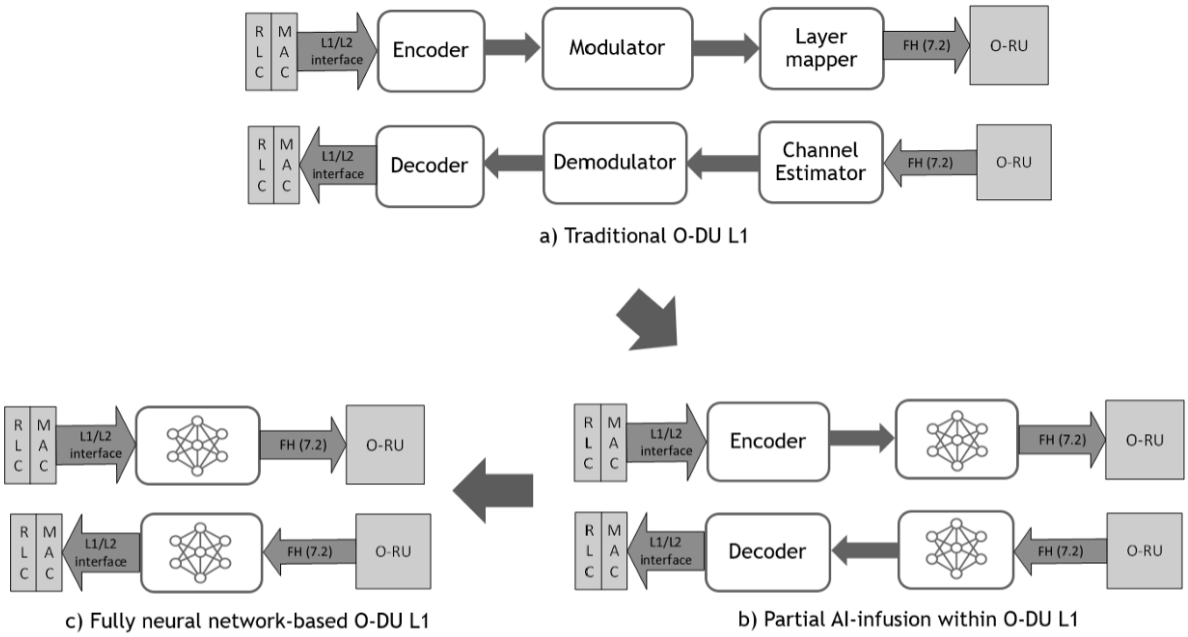

Figure 2.1-1 An example of native AI based RAN, with AI infusion in O-DU node

2.2 Cross-domain AI

Domains in telecommunications networks can exist within boundaries of a physical network or extend beyond physical boundaries into virtual world, also known as digital twin space. To unleash the true potential of AI for wireless, AI-native architecture will extend beyond a domain boundary, with distributed AI paving the way for cross-domain intelligence sharing and optimization. Network functionalities spanning across multiple physical network domains exist today, with examples like centralized management and data analytics service (MDAS) [19] providing cross-domain data analytics service across various physical network domains like RAN, CN, and so on, as depicted in Figure 2.2-1.

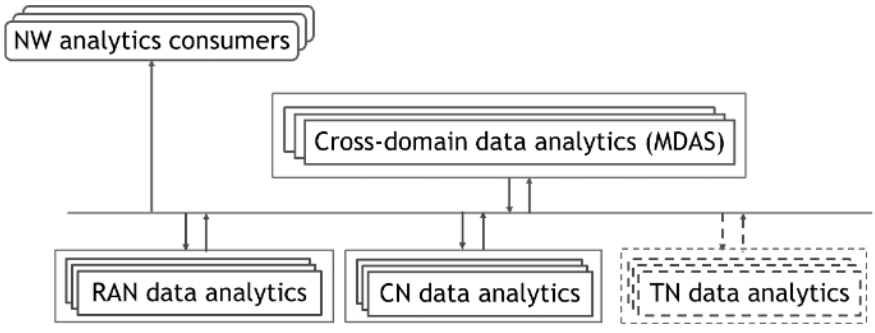

Figure 2.2-1 Centralized management and data analytics service (MDAS)

Going beyond the boundary of physical network space, cross-domain AI will also enable realization of network digital twins, the digital replicas of the physical network entities in the virtual space with acute precision [20]. Cross-domain data collection,

control and optimization loops running between the physical network domain(s) and their digital twin counterparts, as depicted in Figure 2.2-2, can be enriched with enhanced capabilities through cross-domain AI [21].

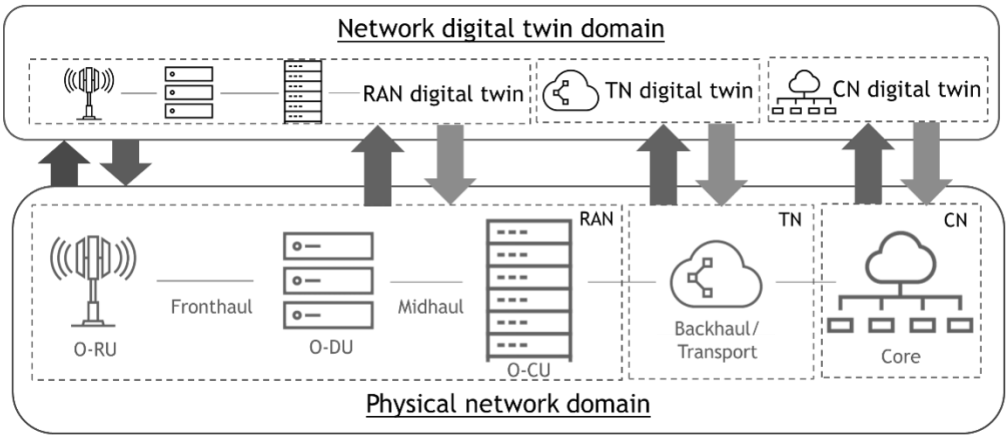

Figure 2.2-2 Cross-domain interaction between physical network domains and network digital twin domains

In a nutshell, the all-encompassing emergence of AI into 6G networking space will permeate across network domains and physical network boundaries, enabling AI collaboration and AI integration across a multitude of network domains in both physical and digital twin spaces.

Cross domain AI

Collaboration and integration of AI-enabled functionalities across different domains

- *Domains can map to network domains (like Radio Access Network, Core Network, Transport Network, Network Applications etc. and network digital twin(s)) or other domains.*

3. AI aspects and their impact on architecture

3.1 Data Ingestion and Management

Data is one of the key elements of artificial intelligence, especially for deep learning models based on neural networks, which require a large amount of training data and test data. In terms of data volume, mobile communication network has natural advantages. All the way from UE to wireless access network, transmission network, and core network, thousands of data fields and indicators will be generated during network operation, involving different software and hardware, functions, and protocol stacks. In general, the more the volume of data, the greater the potential of AI. However, the quality of data ultimately determines the upper limit of AI performance.

Therefore, efficient and feasible data collection and management schemes are crucial for AI-native network.

Centralized data ingestion and management makes it difficult to meet real-time native AI requirements, such as physical layer AI. This entails the need for a data-driven architecture that is able to connect distributed data domains, and distributed, cross-domain, network intelligence fabric as opposed to a centralized one.

As network structures, UE types, UE behaviors, data service requirements, and system resource configurations of 6G will be highly dynamic, time-sensitive, and coupled, data ingestion and management on mobile networks will face many challenges. A major difficulty lies in data ingestion because data is scattered across different devices and at different layers. On one hand, various data types and complex structures make it difficult to analyze data, and highly abstract data hinders understanding. On the other hand, coupled data with complex correlations make data mining even more difficult. Coping with these challenges requires – 1) standard data attribute generalization, 2) efficient clarification of relationships between data, 3) further exploration of in-depth associations between relationships, and 4) clear representation of data and key information inherent in relationships [31].

Currently, the O-RAN architecture supports two flavors of RAN Intelligent Controller (RIC) – Non-Real Time RIC and Near-Real Time RIC, building on the 3GPP architecture, and adds interfaces such as A1, E2, O1, and O2 around these two nodes, providing a centralized data ingestion and management framework for network intelligence. In a centralized data ingestion and management system, data needs to undergo collection, aggregation, and analysis before being sent to the execution unit by RICs. Obviously, this centralized data ingestion and management is difficult to meet real-time native AI requirements, such as physical layer AI. The network requires efficient real-time data ingestion, processing, and storage capabilities to meet Real time native AI requirements. Due to the high openness of the O-RAN architecture, the base station equipment may come from different suppliers. It is therefore beneficial to establish a unified data ingestion model as well as standardized data formats. Besides, networks need to support real-time and reliable data ingestion and transmission, real-time data pool, real-time data computing, and other technologies.

3.2 Distributed Intelligence

3.2.1 Native AI for future wireless network

The 5G communication system can provide a modular framework to meet stringent latency and reliability requirements. However, merely introducing AI technology into a wireless network to solve a certain network optimization problem does not make the network intelligent. The continuously changing radio environment requires retraining and updating of fixed ML models, resulting in repetitive work and hindering the intelligence of the wireless system.

The future communication system is not only expected to apply AI to enhance certain functions such as energy-saving and mobility management, but it is also designed to

integrate AI into every aspect of wireless communication systems to depict the vision of intrinsic AI. Therefore, to bring autonomous learning, decision-making, self-optimization, and self-evolution, an AI-Native radio access network will be an intelligent loop that includes intelligent data perception, intelligent modeling, distributed architecture, and intelligent monitoring.

- Intelligent data perception: GANs can avoid transferring large amounts of data between nodes in the network, generating required data to simulate real data and improving model performance.
- Intelligent Modeling: AutoML technology can automatically select a machine learning model and optimization algorithm, reducing manual intervention for autonomous learning capabilities.
- Distributed Architecture: Distributed AI server architecture uses AI units distributed across network nodes to reduce data transmission load, protect data privacy, and improve model performance.

Intelligent Monitoring: Incorporating human control into the decision-making process can improve AI algorithms' decision-making and help machines understand user preferences, using reinforcement learning to adapt to external environmental changes.

3.2.2 Trustworthy Distributed Intelligence

6G is envisioned to connect human, digital, and physical worlds considering three core values at its center: (i) trustworthiness for 6G as a backbone of society (ii) inclusiveness for 6G to be available for everyone and everywhere, and (iii) sustainability for 6G with regard to environmental, social and economic aspects. To realize such a vision, Distributed Intelligence using Machine Learning (ML) is expected to play an important role as means to automate and optimize complex network operations in the RAN, other network domains and across domains. Therefore, it is of utmost importance to consider trustworthiness of ML in network automation, especially with hundreds to thousands of ML pipelines anticipated to be deployed across 6G end-to-end mobile networks.

Any ML pipeline comprises of at least three different stages: (i) the data stage to collect, validate and preprocess the data to make it suitable for training (ii) the training stage to train, test and validate the ML model and (iii) the inference stage to deploy and monitor the behavior of the ML model. Therefore, the requirements for trustworthiness may vary for different stages of an ML pipeline and as such the related ML trustworthiness needs to be incorporated and managed (i.e., configured, monitored, and measured) by relevant communication stakeholders.

Key technology enablers for trustworthy ML are explainable ML, fair ML, and robust ML; these enablers should be managed in a unified way across distributed ML pipelines, as well as for single-domain and cross-domain interactions. An initial standardization work in this area is the work in ETSI ZSM012 [14] , cf. section 2.3.4.

There, new ML Management Services (MnSs) to support ML-driven end-to-end network and service automation operations are specified. The specification contains MnSs for Management Domain Trust & Security, comprising of ML performance and ML trustworthiness-related management services between ML pipeline orchestrators and ML trust engines of individual management domains and an end-to-end service management domain.

With the continuous development, energy consumption is a major cost for mobile network operators. AI/ML solutions are in general computation expensive in model training. The training cost from the increase of AI/ML solutions places further challenges to the goal of sustainable development of next generation mobile networks. In contrast to centralized learning, which implements AI/ML solutions at a central place (e.g., the cloud) and leverages the network infrastructure to accumulate raw training and feedback data, edge learning facilitates the AI/ML model training using local data and fits into the disaggregated network architecture.

Specifically, distributed AI/ML agents located in different components of the network (including user devices) parallelly coordinate the training of an AI/ML model for a specific service (e.g., inference or decision-making). The local AI/ML model parameters are then shared over the network and aggregated at the edge/cloud, thus avoiding uploading huge amounts of raw data. Consequently, the local AI/ML models at distributed AI/ML agents are adapted to a global model. By splitting an AI/ML model into agent-level and server-level offspring AI/ML models, the communication workload and the computation intensity at a distributed AI/ML agent can be potentially well-balanced.

The wireless connectivity plays a critical role in improving the edge learning performance. Due to the uncertain wireless networking environment (e.g., the random mobile user arrivals/departures, the sporadic wireless service requests), the radio resources have to be appropriately allocated in accordance with the temporal transmission quality variations between the server and the geographically distributed user devices. Conventional single/multi-agent reinforcement learning (RL) is an efficient solution to achieve the optimized resource allocation policy. However, solving an optimal radio resource policy requires online interactions with the mobile networks, which can lead to painful consequences during the exploration phase of RL. This asks for the development of offline RL based resource allocation solutions. To make the resource allocation more robust to the non-stationary network uncertainties, meta-learning and transformer can be adopted to empower offline RL solutions.

3.2.3 Edge Node Enhanced Distributed Intelligence

Distributed intelligence is a key enabler technology for 6G networks. It is a potential approach to improve network performance, enhance security, increase reliability, and support for emerging applications and services. With the progress of virtualization, edge node is proposed for integrating the abilities of RAN, edge computing, network functions of CN and so on. An edge node is a physical or virtual machine located at

the edge of a network. The edge node is an option for enhancing distributed intelligence. Specifically, lots of application or service requests are from the edge of the network, and the distributed intelligence embedded in the edge nodes can process the requests with an intelligence method.

Edge nodes and distributed intelligence are complementary to build an efficient and scalable future network. On the one hand, the edge node provides computing power for AI model training, storage for data, and network connectivity for information exchange. On the other hand, distributed intelligence has trained AI or ML algorithms for application inference. Additionally, distributed intelligence can serve network functions and promote network intelligence.

The edge node also has the advantage of enabling distributed intelligence, as follows. (i) Lower latency: the location of the edge node is closer to the users, which results in lower latency. This is critical for the new applications (e.g., extended reality) of the future network; (ii) Improved accuracy: the edge nodes perform more advanced online data analytics using massive amounts of data, improving the training accuracy of AI or ML models; (iii) Enhanced privacy: the local data processing and analysis ensure data security, especially for sensitive data.

3.2.4 Distributed Intelligence for Native AI

Future 6G network is envisioned to be native AI. Besides optimizing network performance using AI technique (AI for network), 6G network can also provide AI service to any over-the-top (OTT) applications (network for AI), such as AR/VR and autonomous driving. Different types of AI tasks also mean different requirements for the future network. In particular, network optimization across multi-vendor networks and UEs would require close collaboration among different vendors with respect to sharing of training data and models. Besides, some over-the-top applications, such as AR/VR and autonomous driving, would require low latency for AI inference and result delivery.

Compared to conventional centralized intelligence architecture, distributed intelligence architecture has its advantages to address above mentioned requirements. For example, by using Federated Learning technique, different network components (including terminal devices) can train an AI model jointly using their local data, which can achieve the global optimization without leaking any sensitive data to other partners. To fulfill the stringent requirements from the over-the-top applications, network can fully utilize the computation resources within the network across different components, e.g., AI training computing in the cloud and AI inference computing on the edge.

On the other hand, to realize the distributed intelligence architecture, an entity responsible for the AI task management and computation resource management is foreseen as needed. Managing the distributed intelligence in large scale could be a challenging task considering the vast number of network components (including terminal devices) and possible AI applications.

3.2.4.1 Major Enablers and Challenges of AI infused Wireless Networks

Next generation wireless networks will be fully enabled with advance silicon fabric which will give the network sufficient, flexible and software programmable capabilities for high performance compute, the most important tool to infuse AI/ML based Deep Neural Network Functional units supported at the telco and device edge including RFICs, Base Band Processors, Host and Co-host processors.

Architecture and deployment of these networks need highly distributed AI, where some of the requirements like On-Device Edge AI (Distributed Device Inference), Telco and Metro Edge AI (Edge Device inference with Telco and Metro AI modules), Public Edge AI (Federated and Cross Domain AI) will be challenging to support in a wholistic and economical way.

3.2.4.2 AI/ML enabled HPC Platforms

It is important to mention the basic fabric changes to the wireless networks Edge infra to support DNN (Deep Neural Network) functional unit, so that AI/ML functionalities supported to all distributed elements of the network. Infrastructure as a service (IaaS) and Platform as a Service (PaaS).

Open-source SDKs like CUDA, CUDA-X, OpenGL, Helix, ROCm supported by various GPU and coming advance GPU platforms compliant to domain specific architecture are some of the examples of next generation AAL (Abstraction and Acceleration Layer) infrastructure capable of supporting advance wireless solutions with a flexibility of launching multiple slices of entirely different nature.

3.2.4.3 Advancements to the AAL platforms

Advance AAL platforms may support the following features:

- Highly distributed, DNN/ CNN FU enabled Edge Framework.
- The coarse grained reconfigurability to realization of heterogenous, dynamic, and elastic workloads, while the fine-grained reconfigurability will allow for ASIC like optimization for DSP/AI workloads.
- A single fabric should handle all computations, DSP kernels, AI-ML workloads in a dynamic real time elastic manner.
- Performance: Massively Parallel.
- Software Controlled Hardware.
- Billion parallel threads
- High reliability, availability and adaptability
- Major Accelerator Types: Lookaside and Inline. Inline Acceleration has a better future due to heavy dependency of nG Applications on AAL Platforms

3.3 Intent-Driven Network Management

The evolution toward a 6G network architecture is characterized by the increasing adoption of AI technologies. Despite AI's benefits in realizing different use cases, e.g., mobility prediction, network energy saving, and load balancing in RAN [23], user experience prediction, congestion prediction, and traffic prediction for NWDAF in the core network [24], the difficulty in managing AI models, and the data infrastructure needed to keep these models healthy make the human operator's task very complex.

To manage this emerging complexity, boosted by the multitude of different applications that the network will support in the 6G timeframe, increased levels of automation are needed. The idea is that, instead of introducing new manual operations, humans are relieved from the task of managing the network and moved into a supervisory role, where they express requirements via intents fueling zero-touch autonomous network operations.

Intents are defined as “the formal specification of all expectations including, requirements, goals, and constraints given to a technical system” [25]. A human can use intents to specify what to achieve in the network but not how to achieve a desired state. In this sense, intents can be seen as a first-class enabler of AI-driven decision-making for autonomous network operations.

The impact of intent-driven network management has cross-domain implications. Indeed, the network is envisioned to be organized as a set of intent management functions communicating with each other via intent interfaces and fostering autonomous network operations based on intent handling. Hence, intents are not only used by humans but, even more significantly, between machines and subsystems.

Different standardization bodies have already started executing on intents, e.g., APIs and intents common models, where vocabulary and semantics can be used for intent specification in TMForum [26], and the work in 3GPP where intent-driven management services for mobile networks are introduced [27]. O-RAN can leverage the ongoing work and harmonize models and interfaces to foster end-to-end cross-domain integration with external domains, e.g., core network, transport network, cloud infrastructure. In current O-RAN work, WG1 specifies the high-level use of intents with the Non-RT RIC sending intents to Near-RT RIC to drive optimization at the RAN level in terms of expected behavior, and with the Near-RT RIC supporting interpretation and execution of intents [28].

Different research aspects are in the scope of intent-driven network management for O-RAN. Interfaces and model specifications will be most likely driven by the standardization groups already starting from 2023.

Within an O-RAN 6G outlook, different research opportunities come with intents that need further investigation:

- What are the implications of intents on the cross-domain aspect, and what does it mean for O-RAN?

- What are the challenges and key technologies to realize hierarchical intent-based management?
- Conflict detection and resolution among intents, how to deal with multiple intents?
- How to foster the evolution from imperative policy specifications to AI-driven decision-making driven by intents?

To summarize, different research aspects are in the scope of intent-driven network management for O-RAN. Interfaces and model specifications will be most likely driven by the standardization groups already starting from 2023. As a future research outlook, the following aspects can be explored: cross-domain integration, hierarchical intent management, conflict detection and resolution, and AI-based techniques for realizing autonomous system operations.

Future Service Management orchestration framework will provide full support of AI driven Orchestration, since DNN FU (AIaaS) becomes an integral part of IaaS. Additional Modules and Interfaces for Cross Domain AI and Federated AI are also likely to be included.

3.4 Collaboration across Domains

3.4.1 Collaboration across disaggregated RAN functions

Native AI calls for the deep integration of computing and communication to provide end-to-end intelligent services for everyone and everywhere. Currently, both 3GPP and O-RAN design centralized AI for the RAN domain. Specifically, in 3GPP [29], RAN data was collected in OAM or gNB for model training and inference, while ORAN implementing AI through Non-RT RIC and Near-RT RIC. However, plug-in AI makes it difficult to achieve the differentiated requirements of diverse services, and massive data transmission and AI/ML training will further increase the network overhead. Therefore, AI capabilities will be embedded within disaggregated RAN functions, providing a more flexible and customizable AI service through collaboration.

Embedded AI can use local data for AI training, avoiding the overhead and security problems caused by data transmission over the network. Therefore, how to design AI collaboration mechanism across disaggregated RAN functions is critical for native AI. First of all, the enhancement of RIC function should be considered (i.e., the function of AI collaboration control should be provided). Then, a real-time RAN intelligent controller should be designed to achieve real-time control of AI embedded within RAN functions, in addition to Non-RT RIC and Near-RT RIC. Finally, the collaborative methods of the RAN embedded AI components should be further investigated. Multiple AI agents should support distributed learning such as federated learning, split learning, and transfer learning, and well-trained models should be able to share among multiple RAN functions.

3.4.2 RAN-CN convergence and collaboration

Currently, both CN and RAN architectures adopt a centralized plugin AI. Specifically, for the O-RAN architecture, the intelligent control of RAN is realized by Non-RT RIC and Near-RT RIC. The Non-RT RIC supports intelligent control and optimization of RAN by providing policy-based guidance, ML model management and enrichment information to the Near-RT RIC. And the Near-RT RIC enables near real-time control and optimization of E2 nodes functions and resources via fine-grained data collections and actions over the E2 interface. While for the core network, the intelligent network control and optimization are realized by NWDAF, which automatically perceives and analyzes the network based on collected network data and participates in the whole life cycle of network planning, construction, optimization, operation and maintenance. It can be seen from the above that AI functions in the CN and RAN are designed for their respective scenarios and are relatively independent.

Facing the diversified and differentiated requirements of the future 6G network, it is necessary to embed AI functions into the network elements of CN and RAN in a distributed manner and support end-to-end intelligent services through the collaboration of CN and RAN. However, the centralized plug-in AI in the existing RAN-CN separation architecture requires frequent signaling and data exchange between RAN and CN when performing network intelligent collaboration, which makes the collaboration process complex with high resource overhead and delay.

Therefore, in the future 6G network, some CN functions should be sunk to the edge network and converged with the corresponding RAN functions. For example, the NWDAF in the CN can be sunk to the edge node to integrate with the relevant AI functions in the non-RT RIC or even the near-RT RIC, thereby simplifying the intelligent collaboration process between RAN and CN and reducing resource overhead and delay. In addition, since the relevant training and inference processes can be completed independently at the edge nodes, the data does not need to be transmitted to a higher-level central cloud, thus improving privacy and security.

In addition to the converged CN-RAN architecture, the cross-domain AI collaboration between CN and RAN is also important for achieving end-to-end intelligence. First of all, the RAN architecture should be enhanced. SMO needs to provide cross-domain AI management and orchestration of RAN and CN. For example, the MDA (management data analytics) in 3GPP [30] forms a part of the management loop, and it brings intelligence and generates value by processing and analyzing management and network data from different domains. Each domain should design the AI capability exposure function to provide available computing resources, models, and other AI information to other domains. Then, the AI workflow and life cycle should be task-oriented and consider inter-domain collaboration processes such as cross-domain data collection, model sharing, inter-domain distributed learning methods and so on.

4. Conclusions and Summary

This document provides an overview of the landscape associated with the emergence of AI/ML as a vital ingredient of the next generation network. It identifies the key areas of architecture that are impacted by native and cross-domain AI and provides detailed discussion of the expected impact. Key areas such as data management, intent-driven network management, distributed intelligence and cross-domain collaboration are tackled. The document provides a foundation upon which further work can be done in O-RAN nGRG in enabling and accelerating the infusion of AI/ML into the network.

5. References

[1] 6G wireless Native AI architecture and technology white paper.
[2] 3GPP TR 28.908, “Study on Artificial Intelligence / Machine Learning (AI/ML) management”
[3] 3GPP TR 37.817, “Study on enhancement for data collection for NR and EN-DC.”
[4] RP-213599, “Study on artificial intelligence (AI)/machine learning (ML) for NR air interface.”
[5] 3GPP TS 23.288, “Architecture enhancements for 5G system (5GS) to support network data analytics services.”
[6] 3GPP TS 28.533, “Management and orchestration; Architecture framework.”
[7] X. Lin, L. Kundu, C. Dick, and S. Velayutham, “Embracing AI in 5G-Advanced Towards 6G: A Joint 3GPP and O-RAN Perspective,” arXiv preprint arXiv:2209.04987, 2022.
[8] 3GPP TR 28.908, “Study on Artificial Intelligence / Machine Learning (AI/ML) management”
[9] 3GPP TR 33.877, “Study on the security aspects of Artificial Intelligence (AI)/Machine Learning (ML) for the NG-RAN”
[10] O-RAN.WG2.AIML-v01.02.4, O-RAN WG 2, AI/ML workflow description and requirements, February 2021.
[11] CMCC.AO-2022.10.17-WG3-D-WID-AIML_in_O-RAN-v2, October 2022.
[12] ITU-R M.2516-0, “Future technology trends of terrestrial IMT systems towards 2030 and beyond”, Nov 2022
[13] ETSI, Zero Touch Network and Service Management (ZSM), Reference Architecture, ETSI GS 002 1.1.1, August 2019.
[14] ETSI, Zero Touch Network and Service Management (ZSM), Enablers for Artificial Intelligence-based Network and Service Automation, ETSI GS 012 1.1.1, December 2022.
[15] J. Wu, R. Li, X. An, C. Peng, Z. Liu, J. Crowcroft and H. Zhang, “Toward Native Artificial Intelligence in 6G Networks: System Design, Architectures, and Paradigms”, arXiv preprint arXiv:2103.02823, 2021.
[16] S. D’Oro, M. Polese, L. Bonati, H. Cheng and T. Melodia, “dApps: Distributed Applications for Real-Time Inference and Control in O-RAN,” in IEEE Communications Magazine, vol. 60, no. 11, pp. 52-58, November 2022.
[17] L. Velasco et al., “End-to-End Intent-Based Networking,” in IEEE Communications Magazine, vol. 59, no. 10, pp. 106-112, October 2021.
[18] Neural Receiver for OFDM SIMO Systems
[19] 3GPP TS 28.533, “Management and orchestration; Architecture framework.”
[20] ITU-T Recommendation Y.3090, “Digital Twin Network – Requirements and Architecture.”

[21] X. Lin, L. Kundu, C. Dick, E. Obiodu and T. Mostak, “6G Digital Twin Networks: From Theory to Practice”, arXiv preprint arXiv:2212.02032, 2022.
[22] RP-213602: AI/ML RAN enhancem–nts - AI algorithms for network energy saving, load balancing and mobility optimization
[23] RP-213599: “Study on artificial intelligence (AI)/machine learning (ML) for NR air interface.”
[24] SP-220678: Study on Enablers for Network Automation for 5G - Phase 3
[25] TMForum – Intent-based automation - https://www.tmforum.org/opendigitalframework/intent-based-automation/
[26] TMForum – T–290 - https://www.tmforum.org/resources/technical-report/tr290-intent-common-model-v3-0-0/
[27] 3GPP - TS 28.312 – “Management and orchestration; Intent driven management services for mobile networks”
[28] O-RAN.WG1.Use-Cases-Detailed-Specification-v09.00
[29] 3GPP TR 37.817: Study on enhancement for data collection for NR and ENDC.
[30] 3GPP TR 28.809: Study on enhancement of Management Data Analytics (MDA).
[31] 6GANA, “Data Acquisition and Analysis of B5G6G Network Intelligence white paper [R],”2022.
[32] 3GPP TR 21.905: “Vocabulary for 3GPP Specifications”
[33] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
[34] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[35] J. Braams. (2001, Feb.) The Babel package. [Online]. Available:http://www.ctan.org/tex-archive/macros/latex/required/babel/
[36] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathinelevated channel low-temperature poly-Si TFT,” IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[37] F. Delorme et al., “Butt-jointed DBR laser with 15 nm tunability grown in three MOVPE steps,” Electron. Lett., vol. 31, no. 15, pp. 1244–1245, 1995.
[38] R. K. Gupta and S. D. Senturia, “Pull-in time dynamics as a measure of absolute pressure,” in Proc. IEEE International Workshop on Microelectromechanical Systems (MEMS’97), Nagoya, Japan, Jan. 1997, pp.290–294.
[39] B. D. Cullity, Introduction to Magnetic Materials. Reading, MA: Addison-Wesley, 1972.
[40] Y. Okada, K. Dejima, and T. Ohishi, “Analysis and comparison of PM synchronous motor and induction motor type magnetic bearings,” IEEE Trans. Ind. Applicat., vol. 31, pp. 1047–1053, Sept./Oct. 1995.
[41] M. Coates, A. Hero, R. Nowak, and B. Yu, “Internet tomography,” IEEE Signal Processing Mag., May 2002, to be published.
[42] “Defining AI native: A key enabler for advanced intelligent telecom networks”, Ericsson, https://www.ericsson.com/en/reports-and-papers/white-papers/ai-native