O-RAN next Generation Research Group (nGRG)

Research Report

O-RAN Native AI Architecture Description

Report ID: RR-2023-02

Contributors:

CSCN

CTC

VTT

Lenovo

CICT

Release date: 2023.12

Authors

CSCN	Sun Junshuai	13911937519@139.com
CTC	Wang Qingtian	wangqt08@chinatelecom.cn
VTT	Chen Tao	tao.chen@vtt.fi
Lenovo	Yang Shuigen	yangsg3@lenovo.com
CICT	Sun Wanfei	sunwanfei@catt.cn

Reviewers

NVIDIA	Lopamudra Kundu	lkundu@nvidia.com
Nokia	Niraj Nanavaty	niraj.nanavaty@nokia.com
Qualcomm	Douglas Knisely	dknisely@qti.qualcomm.com
Dell	George Ericson	George.Ericson@dell.com

Disclaimer

Copyright

Executive summary

Native Artificial Intelligence (AI) is the enabler technology for 6G. The RAN Intelligence Controller (RIC) of O-RAN is the potential approach for native AI. For novel scenarios in the future, AI can help to improve the quality of user experience and network efficiency. The report analyzes the requirements of Native AI in 6G and discusses the principles and features of native AI, followed by a recommended future framework for native AI.

Table of Contents

1 Definitions and Abbreviations

1.1 Definitions

E2: E2 is a logical interface connecting the NRT-RIC with an E2 Node as defined in [1].

Near-RT RIC: O-RAN near-real-time RAN Intelligent Controller: a logical function that enables near-real-time control and optimization of RAN elements and resources via fine-grained data collection and actions over E2 interface. It may include AI/ML workflow including model training, inference and updates.

Non-RT RIC: O-RAN non-real-time RAN Intelligent Controller: a logical function within SMO that enables non-real-time control and optimization of RAN elements and resources, AI/ML workflow including model training, inference and updates, and policy-based guidance of applications/features in Near-RT RIC.

O1: Interface between management entity and O-RAN managed elements, for operation and management, by which FCAPS management, PNF (Physical Network Function) software management, File management shall be achieved.

O2: Interface between management entities and the O-Cloud for supporting O-RAN virtual network functions.

A1: Interface between Non-RT-RIC and the Near-RT RIC functions.

1.2 Abbreviations

3GPP	3rd Generation Partnership Project
5GC	5G Core
AI	Artificial Intelligence
CN	Core Network
ML	Machine Learning
NRT-RIC	Near Real Time Radio Intelligence Controller
nRT-RIC	non Real Time Radio Intelligence Controller
RAN	Radio Access Network
RIC	Radio Intelligence Controller
RRM	Radio Resource Management
SMO	Service Management and Orchestration
TN	Transport Network

2 The Architecture of O-RAN: Overview

This chapter refers to O-RAN specification 0, and describes the standardized architecture of O-RAN, that serves as an important baseline for the research of the future RAN.

There are three parts as the basic characteristics of O-RAN that are different from RAN defined by 3GPP, etc. NRT-RIC, nRT-RIC, and O-Cloud & SMO (Service Management and Orchestration) are shown in Figure 1.

2.1 Overall Architecture of O-RAN

Figure 1 provides an overview of the O-RAN architecture. The SMO utilizes four key interfaces to connect to O-RAN network functions and to the O-Cloud. The interfaces are the A1, O1, O2, and Open-Fronthaul-M-Plane. Figure 1 also illustrates that the O-RAN network functions can be VNFs (Virtualized Network Functions), i.e., VMs (Virtual Machines) or Containers, sitting above the O-Cloud and/or PNFs (Physical Network Functions) utilizing customized hardware. All O-RAN network functions are expected to support the O1 interface for management by the SMO.

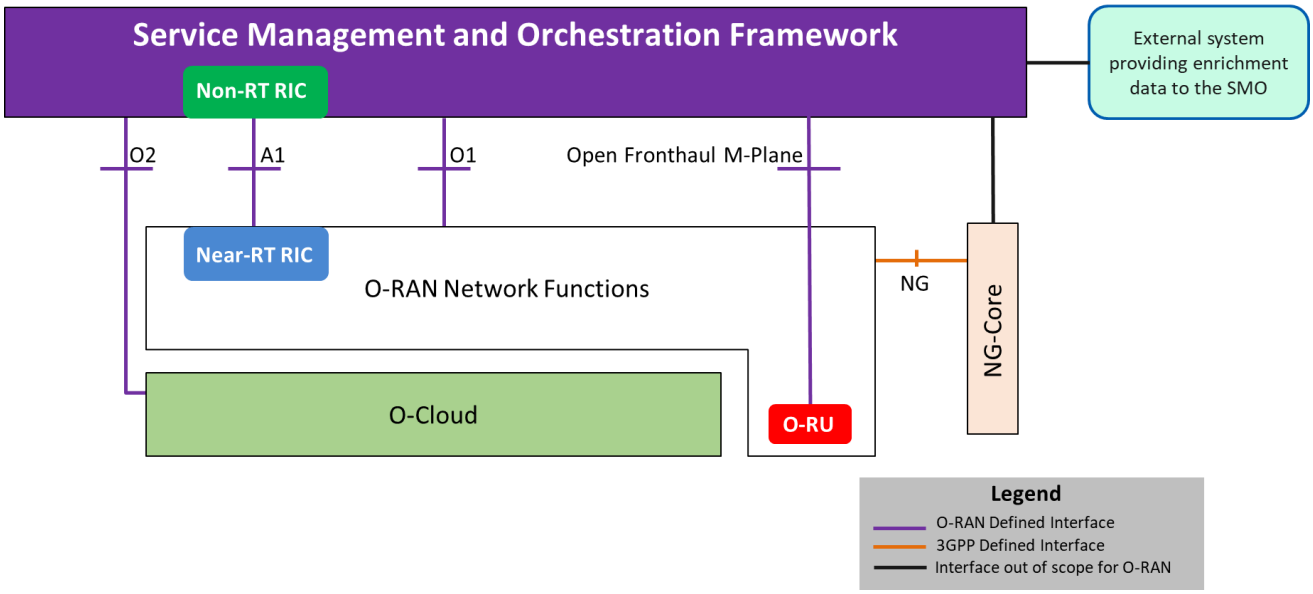

Figure 1: High-Level Architecture of O-RAN

2.2 Non-RT RIC

Non-Real Time RAN Intelligent Controller (Non-RT RIC or nRT-RIC) is the functionality internal to the SMO in O-RAN architecture that provides the A1 interface to the Near-Real Time RIC (NRT-RIC).

The primary achievement of nRT-RIC is to support intelligent RAN optimization by providing policy-based guidance, ML (Machine Learning) model management and enrichment information to the NRT-RIC function so that the RAN can be optimized. It

can also intelligently configure radio resource management function in non-real-time interval (i.e., greater than 1 second).

nRT-RIC can use data analytics and AI/ML training/inference to determine the RAN optimization actions for which it can leverage SMO services such as data collection and provisioning services of the O-RAN nodes.

2.3 Near-RT RIC

Near-Real Time RAN Intelligent Controller (Near-RT RIC or NRT-RIC) is a logical function that enables near real-time control and optimization of E2 nodes' functions and resources via fine-grained data collection and actions over the E2 interface with control loops in the order of 10 ms-1s. The NRT-RIC hosts one or more xApps via E2 interface to collect near real-time information (e.g., on a per UE basis or a per cell basis) and provide value-added services. The NRT-RIC's control over the E2 nodes, and the SMO manages each NRT-RIC via its A1 interface.

The RRM functional allocation between the NRT-RIC and the E2 node is subject to the capability of the E2 node exposed over the E2 interface by means of the E2 Service Model in order to support the use cases described in reference [2]. The E2 Service Model describes the functions in the E2 node which may be controlled by the NRT-RIC and the related procedures, thus defining a function-specific RRM split between the E2 node and the NRT-RIC. For a function exposed in the E2 Service Model, the nRT-RIC may e.g., monitor, suspend/stop, override or control the behavior of the E2 node via policies.

In the event of a NRT-RIC failure, the E2 Node will be able to provide services but there may be an outage of certain value-added services that may only be provided using the NRT-RIC.

2.4 E2 Interface

E2 is a logical interface connecting the NRT-RIC with the DU, CU-CP, CP-UP, and eNB RAN components that support the E2 interface.

•An E2 Node is connected to only one NRT-RIC.

•An NRT-RIC can be connected to multiple E2 Nodes.

The protocols over E2 interface are based exclusively on Control Plane protocols. The E2 functions are grouped into the following categories:

•NRT-RIC Services (REPORT, INSERT, CONTROL and POLICY, as described in [2]).

• NRT-RIC support functions, which include E2 Interface Management (E2 Setup, E2 Reset, Reporting of General Error Situations) and NRT-RIC Service Update (i.e., capability exchange related to the list of E2 Node functions exposed over E2).

3 Requirements of Native AI Architecture of O-RAN

3.1 The Concept of Native AI

Native AI refers to the concept of inherent and built-in AI capability, available or exposed for network or services. Native AI does not just represent a shift in how AI is implemented, but also signifies a profound restructuring of the AI mechanism itself.

Different scenarios require different AI mechanisms. If there are loose causal correlations between different parts of the system, the federated learning [3] model of artificial intelligence should be adopted. On the other hand, if there are clear causal relationships between different parts of the system, the split mode of AI can be utilized. This split AI mode is particularly effective for the intelligent function of RAN and allows for native AI capabilities within the RAN.

Hence, the concept of native AI involves splitting the entire AI system that serves the RAN into multiple subsystems or components based on the specific objectives of the service. Each component is then integrated into the service function of the RAN, to provide a cohesive system. The split AI approach can utilize a distributed architecture where different parts of the system handle model training, derive inference, and carry out data processing, under the centralized control or orchestration. The distributed AI components can be implemented on both the network and the UE, depending on the specific requirements for interaction between the RAN and the UE.

3.2 The Requirements of Native AI in 6G

Native AI is one of the key characteristics under discussion for 6G. In order to enable native AI in 6G, a number of elements, including computing power, data, AI algorithms, and functionalities of RAN, TN, CN and Management should be taken into consideration.

Computing power is the core element of native AI, and 6G applications and services are expected to increase the demand for robust and accelerated computing resources to meet the requirements of these applications and services. The advancement of computing capabilities also plays a vital role in propelling AI development. Powerful computing resource can significantly enhance the accuracy and adequacy of AI model training while reducing time costs. In order to accommodate the evolving scenarios and network functionalities of 6G networks, the development of novel computing architectures and infrastructures for handling extensive data volumes and intricate algorithms is required. This can entail specialized hardware, including graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs), NPUs (Neural Processing Units), data processing units (DPUs). It can also include implementing distributed computing architectures capable of managing the vast quantities of data generated by 6G networks and applications.

Data is another crucial element of native AI in 6G. AI relies on large volumes of data to learn and improve its performance over time, and 6G networks will need to provide a reliable and secure platform for collecting, storing, processing and

retrieving this data. This may involve the development of new data storage and processing technologies, such as edge computing [4] and cloud computing [5]. They can handle the large volumes of data required as well as generated by AI applications in real time.

AI Algorithms are also a key component of native AI in 6G. AI algorithms are the underlying code that allows AI systems to learn from data and make decisions based on that learning. Development of new algorithms may be necessary to handle the unique challenges presented by wireless communication networks, such as variable signal strength, latency, and bandwidth constraints. This may involve the use of machine learning techniques, such as deep learning and reinforcement learning, that can adapt to changing network conditions and optimize performance over time.

Finally, the functionalities of 6G RAN are a critical element of native AI. 6G networks will need to be designed to support the real-time processing and analysis of large volumes of data, as well as the seamless integration of AI applications with existing network infrastructure.

3.3 Programmable RAN toward Native AI

In the context of 6G, the integration of native AI needs an efficient and convenient approach to incorporate AI elements seamlessly. Programmable RAN [6] emerges as a promising solution to drive the development of native AI. By leveraging interfaces like E2SM-KPM [1], programmable RAN enables the collection and interaction of data functions from RAN. This collected data can then be processed and analyzed to uncover meaningful correlations, thereby enhancing the network's overall capabilities. Furthermore, programmable RAN facilitates the adoption of relevant RAN frameworks, enabling the import and update of AI models, ultimately catering to the diverse requirements of various scenarios.

Programmable RAN encompasses three key components: Parameter, Data, and Behavior. Programmable parameter facilitates the seamless adaptation of parameters between AI models and the software-defined RAN functions through a programmable framework and open interface. Programmable data involves the construction of data sets for AI model training and the exploration of data relationships within the RAN functions. Additionally, data can be securely provided to third parties through relevant secured methods. Programmable behavior allows for the modification of RAN actions through the utilization of different AI models. To facilitate this process, a robust framework is required to deploy and manage these models effectively. The framework should encompass a comprehensive set of programmable interfaces and function modules, enabling seamless integration and operation. Furthermore, the programmable behavior aspect empowers the RAN to dynamically update and replace AI models as needed, ensuring flexibility and adaptability in response to changing network requirements.

To enable the implementation of programmable RAN, it is essential to progressively open he traditionally closed protocol stack within the RAN. This involves enhancing the functionality at the protocol stack level and standardizing and generalizing the newly opened interfaces. In the context of the ongoing evolution of native AI,

programmable RAN catalyzes for advancing toward a more open and intelligent RAN vision. By embracing programmability, the RAN can effectively adapt to dynamic network requirements, foster innovation, and leverage the full potential of native AI.

3.4 Centralized and Distributed AI

For 6G networks, it is anticipated that numerous scenarios or applications could be created anywhere in the network and at any given or arbitrary time. Native AI, as the critical enabler of 6G, should handle those requirements with quick response. Therefore, native AI will manifest in two ways in the 6G:, centralized, and distributed AI.

Centralized AI has a central system that can collect all kinds of required information and data in devices or clouds. It has abundant and powerful computing resources for data processing, AI model training and inferring, and so on. However, centralized AI is unable to process the latency sensitive services or applications requiring quick response. In that context, distributed AI is a great complement to centralized AI.

Distributed AI [7] can be deployed anywhere on a 6G network, e.g., radio access network, or core network. Distributed AI is not only efficient with data privacy and bandwidth, but also is responsive to the application’s actions. Distributed AI has the important characteristic of being either data-driven or application-driven. This is because the cached AI models always serve as regional applications. Centralized AI can help distributed AI to train the complex AI models offline.

Both centralized and distributed AI have their advantages. Centralized AI controls data and resources, making it suitable for applications where data centralization or centralized decision-making brings significant benefits. Distributed AI, on the other hand, leverages the computing power of multiple network elements or devices, facilitating collaboration while addressing privacy concerns. The choice among centralized AI, distributed AI and coexistence of both depends on the specific requirements of the applications, data distribution, and privacy.

4 General Principles and Features of Native AI Architecture of O-RAN

4.1 General Principles of O-RAN Native AI Architecture

The general principles refer to the conditions and objectives that should be considered in designing O-RAN native AI architecture.

- The O-RAN native AI architecture should be based on the architecture of O-RAN as the starting point.
- The final native AI architecture may not converge into one and there can be several potential solutions that describe options to support native AI into O-RAN.

4.2 Technical Features of O-RAN Native AI Architecture

To achieve the target of native AI in the architecture of O-RAN, the architecture should have the following technical features:

- Layered Distributed AI: In the future network, the AI function will be flexibly planned and deployed in the O-RAN logical nodes, rather than being largely centralized. The O-RAN native AI architecture is also required to support a more layered distributed AI function.
- Cross-domain AI collaboration: in the 6G timeframe, AI functions are likely to exist not only in the RAN domain, but also in the network domains like core network, UE, etc. To support this, the O-RAN native AI architecture needs to have the ability to carry out the AI task collaboratively among different logical nodes and across different domains, such as RAN domain and UE domain, or RAN domain and core network domain.
- Guaranteed AI service: In the 6G timeframe, AI may be a service provided to customers (i.e., user equipment, network element) in the O-RAN system, i.e., AIaaS, and the service quality should be guaranteed. Therefore, the O-RAN native AI architecture should have the ability to provide AI service and guarantee the quality of the service.

5 Native AI Architecture of O-RAN

In the 4G/5G system, the incorporation of AI is achieved as a potential add-on, which means AI is deployed with the served network functionalities and is outside of network elements logically. The evaluation method for AI model training should be provided from RAN, and the trained AI models or policies would be sent directly to RAN. There are several unavoidable challenges to this approach.

Firstly, a huge amount of real-time and finer-granularity metrics should be sent out from RAN to add-on AI entity which may have a large capacity impact on the transport network and RAN. This brings not only the high cost of RAN, but also adds to the challenges of interoperability among different vendors.

Secondly, effective inference deeply depends on effective metrics. Otherwise, the policies or commands inferred by the AI algorithm are not available to RAN. Therefore, the full potential of AI network cannot be realized.

In response to the above problems, a new native AI architecture should be proposed. In the native AI architecture, huge amounts of metrics should be avoided to be sent outside of RAN, and metrics for training and inference of AI should be converged into the control flow and data flow of RAN. In this research report, a solution using distributed native AI and digital twin network (DTN) [8] is studied in Figure 2.

In the solution, distributed AI system and distributed DT (digital twin) system are defined. In this way, the functionalities of AI and DT are converged into functionalities of CN, TN, RAN and UE respectively. The distributed system of AI and DT is composed of centralized AI/DT part and distributed AI/DT part. The centralized AI part configures and controls the distributed AI part, and Distributed AI part reports measurements to the centralized AI part. Distributed AI part is converged into serving functionalities of CN, TN, RAN and UE, and accelerates them. The instance of AI algorithms is decided by AI serving objectives which means different serving scenarios with different AI algorithms. Distributed DT system is adjoined to distributed AI system in order to bring offline or online simulation to training or being inferred of AI. Distributed AI part and distributed DT part locally process the metrics of RAN so that ZMR (Zero Measurement Report) is achieved. Distributed AI system and distributed DT system run on the cloud platform (O-Cloud) and leverage computing power from it.

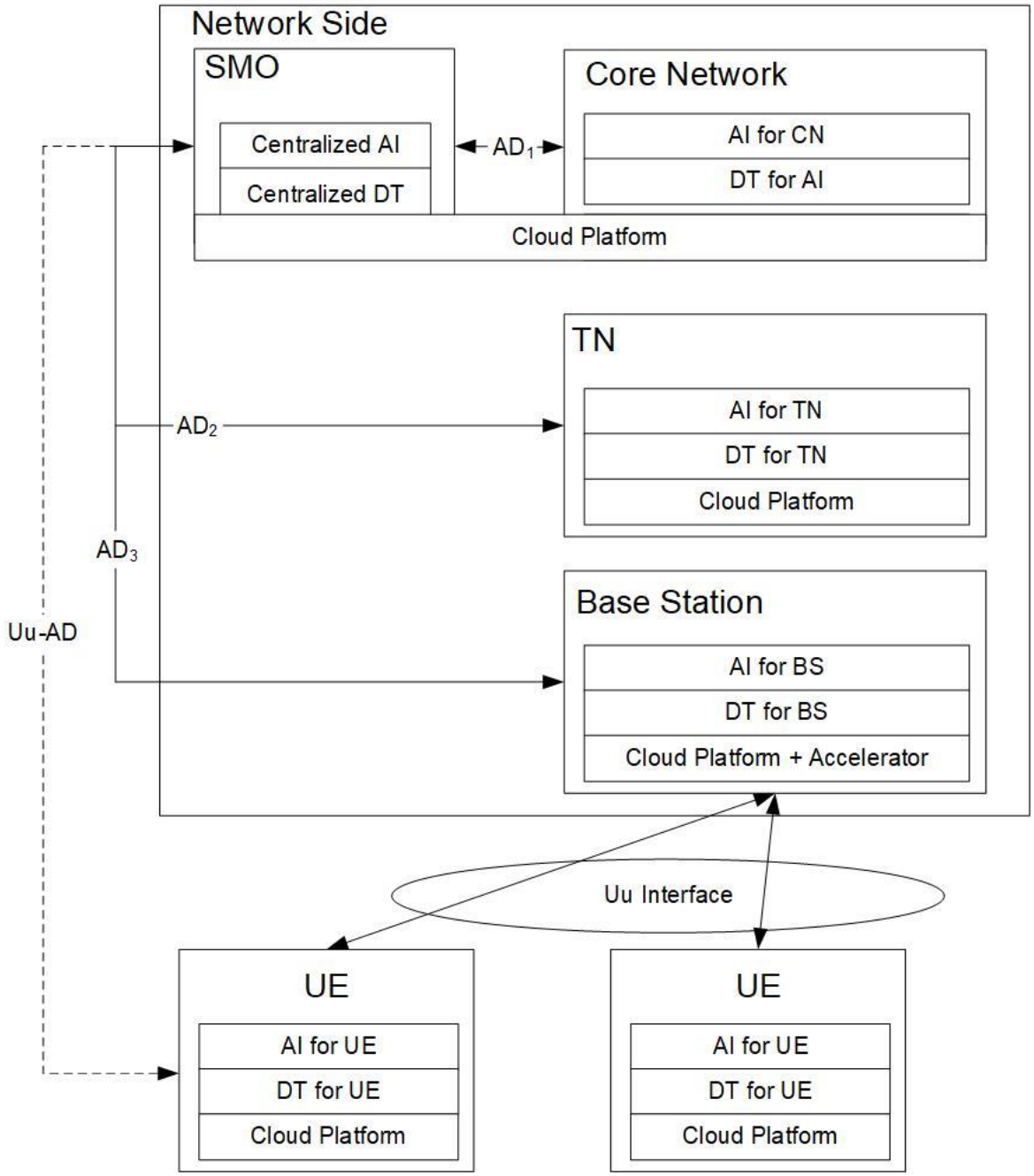

Figure 2: Native AI Architecture of RAN

As shown in Figure 2, there are AI/DT functionalities in CN/TN/RAN and UE respectively.

Centralized AI and DT run on SMO platform and configure the system end to end which means centralized AI and DT configure large-scale models or policies.

For configuring UE, the signaling is carried over the Uu interface [9], which may be considered for 6G specification by 3GPP.

5.1 Collaboration control

The proposed 6G native AI network will provide intelligence, data, computing, management, and other services, e.g., connection services. Enabling this wide range of services requires addressing a diverse set of requirements, which is difficult to meet with traditional wireless networks. Especially, collaboration control between different networks is a new challenge.

Firstly, for differentiated services in different scenarios, the collaboration control function located in the centralized AI/DT will decompose the intent into service requirements on evolved network domains, e.g., UE, radio access network, transport network and core network. The service requirements will affect connection requirements, AI algorithm requirements, data requirements, and computing requirements. The decomposed service requirements are sent to the corresponding collaboration control function in the network domain.

Based on the service requirements from the collaboration control function located in the centralized AI/DT, the collaboration control function located in the network domain will decompose the service requirements into the network function requirements, connection requirements and resource requirements, in order to establish associations and allocate resources. In addition, the network functions, connection, and resources are adjusted adaptively according to the service changes.

Secondly, the collaboration control function will be deployed closer to the network function. Therefore, it will have more real-time management capabilities. For the services with high real-time requirements and low complexity, it will be processed by the collaboration control function located in the UE, radio access network, transport network or core network. For the services with low real-time requirements, larger areas and high complexity, it will be processed by the collaboration control function located in the centralized AI/DT (cloud platform).

5.2 Centralized AI and DT

Native AI can be an important characteristic of 6G, and it can drive 6G network intelligently. The power of AI can be used not only to accelerate the network functionalities and optimize radio resources, but also to run the entire network including CN, TN and RAN as an overall intelligent system.

DT brings the infrastructure to run native AI in 6G. The computing and processing of AI can be run into digital twin domain as online as well as simulation.

The AI and DT systems of the proposed 6G architecture consist of the environment for OAM (orchestration and management) [2], acceleration of algorithms, and flexible routing etc. encompassing an end-to-end system including UE side and network side.

The AI/DT system is a distributed system, of which the centralized AI and the centralized DT are the basic components, and the distributed AI and distributed DT are complimentary components to serve objects on-demand in a distributed manner.

Centralized AI and centralized DT are an organic whole.

Centralized AI entities run for:

a) orchestrating, managing, deploying and controlling all the distributed AI entities.

b) executing functions of centralized AI entities such as training models and processing data with varying time scales (such as 10ms, 100ms, 1m, etc. or some arbitrary periods), making policies or decisions etc.

c) supporting other requirements.

Centralized DT entities run for:

a) orchestrating, managing, deploying and controlling all the distributed DT entities.

b) executing functions of centralized DT entities such as supporting centralized AI to verify the accuracy and validity of policies and commands it produces.

c) producing, managing and running the digital twin entities that centralized AI served.

Centralized AI and centralized DT are implemented in the network side and running on a cloud platform.

5.3 Distributed AI and DT

Distributed AI and distributed DT are the complimentary entities of the AI system and DT system besides centralized AI and centralized DT.

Distributed AI entities run for:

a) receiving and executing commands from centralized AI entities.

b) serving functions of the network.

c) processing measurement reports and metrics locally.

Distributed DT entities run for:

d) receiving and executing commands from centralized DT entities.

e) serving distributed AI entities.

f) providing online simulation for distributed AI entities.

Distributed AI and distributed DT run on CN, TN, BS and UE respectively on demand.

6 Conclusion

The report provides an introduction to a proposed native AI architecture based on the O-RAN architecture. It begins with an overview of the architecture in section 2, followed by the presentation of requirements in section 3. In section 4, the principles and features of the architecture are proposed. Additionally, the report delves into the details of the proposed native AI architecture in section 5.

References

[1] O-RAN E2 General Aspects and Principles (E2GAP) 4.01. Available :

https://orandownloadsweb.azurewebsites.net/specifications

[2] O-RAN Architecture Description 10.0. Available: https://orandownloadsweb.azurewebsites.net/specifications

[3] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE signal processing magazine, 2020, 37(3): 50-60.

[4] Al-Ansi A, Al-Ansi A M, Muthanna A, et al. Survey on intelligence edge computing in 6G: Characteristics, challenges, potential use cases, and market drivers[J]. Future Internet, 2021, 13(5): 118.

[5] Tomkos I, Klonidis D, Pikasis E, et al. Toward the 6G network era: Opportunities and challenges[J]. IT Professional, 2020, 22(1): 34-38.

[6] Bonati L, Polese M, D'Oro S, et al. Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead[J]. Computer Networks, 2020, 182: 107516.

[7] Cao L. Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desci[J]. IEEE Intelligent Systems, 2022, 37(3): 6-19.

[8] Groshev M, Guimarães C, Martín-Pérez J, et al. Toward intelligent cyber-physical systems: Digital twin meets artificial intelligence[J]. IEEE Communications Magazine, 2021, 59(8): 14-20.

[9] NR; NR and NG-RAN Overall Description; Stage 2 (Release 17), 3GPP TS 38.300 V17.6.0, 2023.