O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Research Report on Cross-domain AI

Report ID: RR-2024-02

Contributors:

China Telecom

AsiaInfo

Dell

Lenovo

Nokia

Ericsson

China Unicom

Release date: 2024.02

Authors

Author	Affiliation
Zexu Li	China Telecom (Editor-in-Chief)
Zhanwu Li	AsiaInfo
Hoda Dehghan	Dell
Mingzeng Dai	Lenovo
Niraj Nanavaty	Nokia
Mirko D’Angelo	Ericsson
Tingting Liang	China Unicom

Reviewers

Reviewers	Affiliation
Zhanwu Li	AsiaInfo
Mirko D’Angelo	Ericsson
Niraj Nanavaty	Nokia
Geetha Rajendran	Qualcomm
Ravi Sinha	Reliance Jio
Tao Chen	VTT
Gerd Zimmermann	Deutsche Telekom AG
Lopamudra Kundu	NVIDIA
Hank Kafka	O-RAN
Bernard Guarino	O-RAN

Disclaimer

The content of this document reflects the view of the authors listed above. It does not reflect the views of the O-RAN ALLIANCE as a community. The materials and information included in this document have been prepared or assembled by the above-mentioned authors, and are intended for informational purposes only. The above-mentioned authors shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of this document subject to any liability which is mandatory due to applicable law. The information in this document is provided ‘as is,' and no guarantee or warranty is given that the information is fit for any particular purpose.

Copyright

Executive summary

This document discusses the use cases, requirements, and potential technological directions of cross-domain AI for the next generation networks. Firstly, it provides a brief overview of the current research and application status of Artificial Intelligence (AI) in the various domains of Network as a Service. The report describes the progress of AI-related research within different standards organizations such as 3GPP, O-RAN, ONAP, ETSI, etc. Then this research report further explores cross-domain AI technology. To that end, it focuses on the next-generation network, provides potential technical considerations, and impacts of cross-domain AI on the current network, and presents potential technological directions for the collaboration of data, computing power, and models. This report identifies key research areas in cross-domain AI research and serves as a starting point for further exploration in each key direction.

Table of Contents

Authors 2

Viewers 2

Disclaimer 2

Copyright 2

Executive summary 3

List of abbreviations 6

List of figures 7

List of tables 7

1 .... **Introduction** 7

1.1 Network as a Service domains 8

1.1.1 AI in RAN domain 8

1.1.2 AI in CN domain 10

1.2 Management domain 11

1.2.1 3GPP SA5 Management Data Analytics (MDA) 11

1.2.2 3GPP SA5 AI/ML Management 12

1.2.3 ETSI ZSM AI Enablers 15

1.2.4 ONAP AI Management Practices .... **Error! Bookmark not defined.**

2 .... **Use cases and requirements of cross-domain AI** 16

2.1 Use cases 18

2.1.1 Cross-domain data analysis 18

2.1.2 Intent based E2E smart slicing (ONAP) 19

2.1.3 Digital Twins 21

2.1.4 Cross-domain QoE estimation 26

2.2 Requirements of cross-domain AI 26

3 .... **Potential Solutions for Cross-Domain AI** 27

3.1 Impact of network architecture. 27

3.1.1 Converged RAN-CN architecture 28

3.1.2 Enhanced SMO for cross-domain AI 29

3.2 AI elements collaboration across different network domains 30

3.2.1 Data 30

3.2.2 Algorithm/model 31

3.2.3 Computing resource 36

4 .... **Conclusion** 38

References 39

List of abbreviations

3GPP	3rd Generation Partnership Project
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AnLF	Analytics logical function
CHO	Conditional Handover
CN	Core nETWORK
CSI	Channel State Information
DCAE	Data collection, analytics, and events
DNN	Deep Neural Network
ETSI	European Telecommunications Standards Institute
FL	Federated learning
IP	Intellectual property
KPIs	Key performance indicators
LLM	Large Language Model.
MDA	Management data analytics
MDAS	Management Data Analytics Services
MDs	Management domains
MDT	Minimization Drive Test
MEC	Multi-access edge computing
ML	Machine learning
MnS	Management Services
MS	Micro-services
MTLF	Model training logical function
NDT	Network digital twins
Next G	Next generation
NR	New radio
NWDAF	Network data analytics function
OAM	Operation Administration and Maintenance
ONAP	Open Network Automation Platform
OOM	ONAP operation manager
PLMN	Public Land Mobile Network
QoE	Quality of Experience
QoS	Quality of service
RAN	Radio access network
RIC	RAN Inteligent Controller
RL	Reinforcement learning
RRC	Radio Resource Control
SI	Study Item
SLA	Service Level Agreement
SLS	Service level specification
TAC	Type allocation code
WI	Work Item
Y1	RAN Analytics Exposure Interface
ZSM	Zero Touch Network & Service Management

List of figures

Figure 1.1 Functional framework for RAN intelligence (RAN3 [2]) 8
Figure 1.2 Overview of Near-RT RIC architecture [4].......... 9
Figure 1.3 Logical architecture of O-RAN [5].......... 10
Figure 1.4 Reference architecture for CN data analytics [6].......... 11
Figure 1.5 Coordination between NWDAF, gNB and MDAS producer [7].......... 12
Figure 1.6 AI/ML operational workflow 13
Figure 1.7 RAN domain-specific management function for AI [10].......... 15
Figure 1.8 ETSI ZSM architectural framework [11] 15
Figure 1.9 ETSI ZSM AI enabling areas [12].......... 16
Figure 1.10 ML MS enhancement on intelligent slicing [13] 17
Figure 2.1 A closed-loop structure for E2E slicing [15].......... 20
Figure 3.1 Potential solutions and architecture for cross-domain AI 28
Figure 3.2 An example of a cross-domain data processing process 30
Figure 3.3 An example of split AI/ML inference.......... 34
Figure 3.4 State assisted compute-communication co-design 36
Figure 3.5 Matching of computing resources across different domains 38

List of tables

Table 2.1 Cross-Domain Data analysis for mobility management.......... 18
Table 2.2 ONAP microservices which is related to the E2E slicing use cases.......... 20

1 Introduction

The technical report of O-RAN nGRG-RR-2023-03 (Native and Cross-domain AI: State of the art and future outlook) has already provided the definition of cross-domain Artificial Intelligence (AI) and the potential impact it may have on the network [1]. This research report focuses further on use cases, requirements, and potential technical directions of cross-domain AI. Cross-domain AI means collaboration and integration of AI-enabled functionalities across different domains. These functionalities can map to network domains, (like Radio Access Network (RAN), Core Network (CN), Transport Network (TN), network applications etc. and Network Digital Twin (NDT)) and other domains (like management system, User Equipment (UE), etc.). This report first provides a comprehensive survey of the current state of AI adoption in the domains of RAN, CN, and management system. The report then analyzes the scenarios in which cross-domain AI may be applied, as well as the technical requirements and considerations. Finally, a reference architecture for cross-domain intelligent collaboration is proposed, and possible technical solutions to achieve cross-domain AI collaboration are given.

1.1 Network domains

AI has been widely used in the network domains to solve problems such as network performance optimization, traffic prediction, and assisted decision making. This section introduces the current research status of AI from both RAN domain and CN domain.

1.1.1 AI in RAN domain

(1) 3GPP RAN Intelligence

For the RAN domain, Third Generation Partnership Project (3GPP) RAN3 has studied high-level principles and functional architectures for AI [2] in Release 17 (Rel-17). To support different AI scenarios, 3GPP has proposed several basic functional blocks, which are given in Figure 1.1, including data collection, model training, model inference, and actor. The study focused on three use cases: network energy saving, load balancing, and mobility optimization. AI/Machine Learning (ML) model training and inference can be deployed at Operation Administration and Maintenance (OAM) or 5G Node B (gNB) to provide intelligent capability.

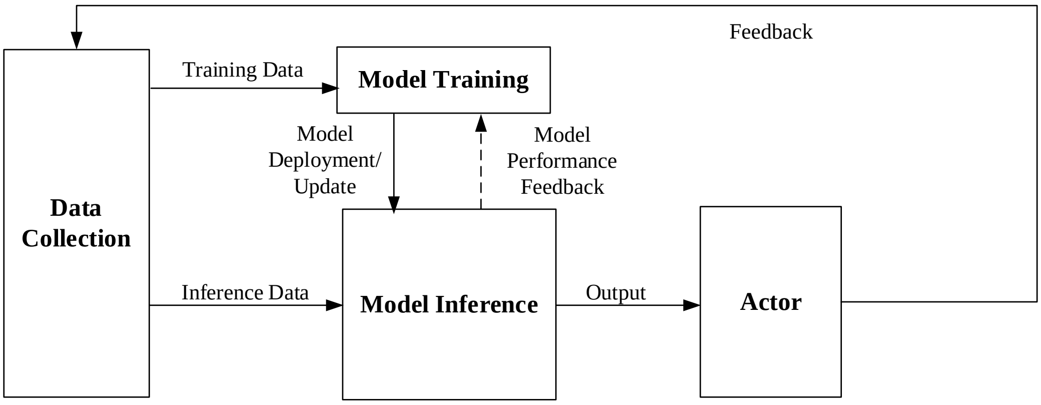

Figure 1.1 Functional framework for RAN intelligence (RAN3 [2])

The normative work based on the conclusion of 3GPP Rel-17 continued in Rel-18, to further specify these enhancements, with a focus on developing standardized interfaces and protocols that enable seamless integration of AI/ML-based solutions into the existing NG-RAN architecture. Besides, 3GPP RAN1 is carrying out a study on AI/ML for 5G New Radio (NR) air interface [3]. The study aims to explore the potential benefits of using AI/ML techniques to optimize the performance of the air interface, such as improving energy efficiency, spectral efficiency, and Quality of Service (QoS), with a focus on three use cases: Channel State Information (CSI) feedback, beam management, and positioning. In Release 19 (Rel-19), there is one new Work Item (WI) and two new Study Items (SIs) on RAN-related AI/ML topics. RAN1 has triggered new work item on AI/ML for NR air interface, which aims to provide specification support for the use cases of beam management and positioning accuracy enhancements. RAN2 has triggered the new SI on AI/ML for mobility in NR, which focuses on mobility enhancement in RRC_CONNECTED mode over air interface. RAN3 triggers the new SI on enhancements for AI/ML for NG-RAN to further investigate new AI/ML based use cases and identify enhancements to support AI/ML functionality.

O-RAN

In O-RAN, Non-real time (RT) and Near-RT Ran Intelligent Controllers (RICs) have been standardized. Non-RT RIC enables long time-scale control and optimization of RAN elements and resources, AI/ML workflow, including model training and updates, and policy-based guidance of applications/features in Near-RT RIC. **Error! Reference source not found.** shows the Near-RT RIC Architecture [4], where Near-RT RIC enables Near-RT control and optimization of RAN elements and resources via fine-grained (e.g. UE level, cell level) data collection and actions over the E2 interface. Besides, Y1 interface from Near-RT RIC to Y1 consumers can be used to expose RAN analytics information, where Y1 consumer can be entities within or outside of the Public Land Mobile Network (PLMN) trust domain that consumes the Y1 services produced by the Near-RT RIC.

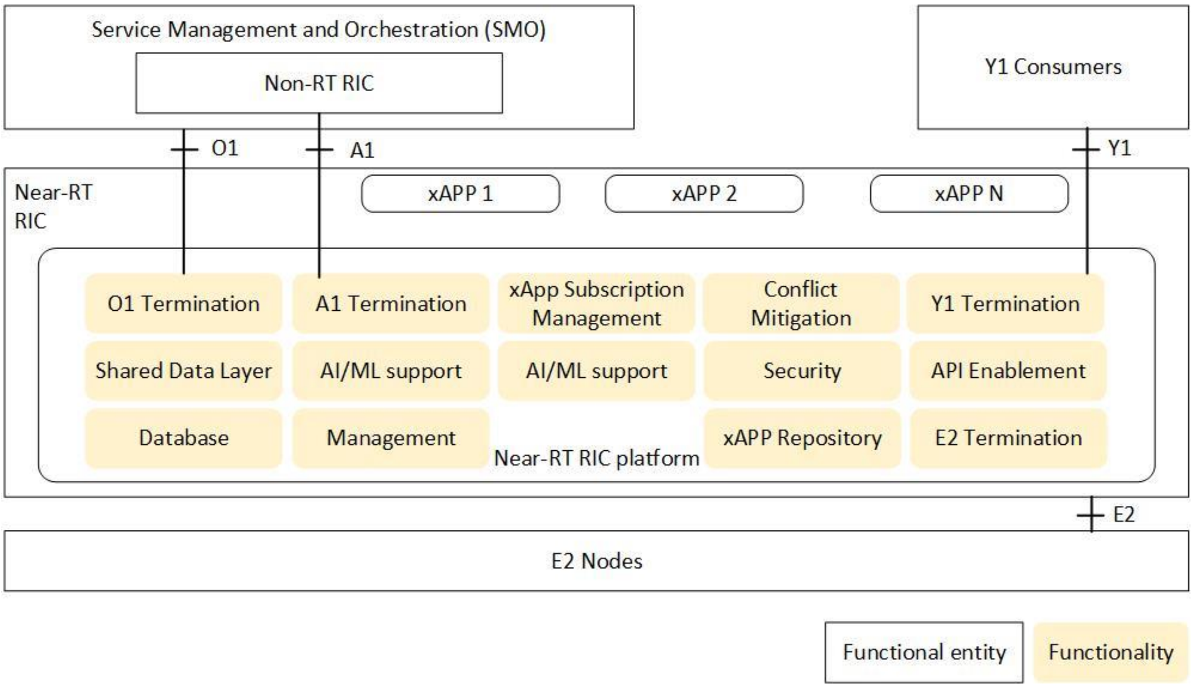

Figure 1.2 Overview of Near-RT RIC architecture [4]

Figure 1.3**Error! Reference source not found.** shows the logical architecture of O-RAN [5]. AI/ML related functionalities can be mapped into three loops. The location of the ML model training and the ML model inference for a RAN function depends on the computation complexity, the availability and quantity of data to be exchanged, and the response time requirements and the type of ML model. O-RAN has set up deeper capabilities to support AI inside the RAN domain. Current O-RAN architecture focuses primarily on AI capabilities on the RAN side.

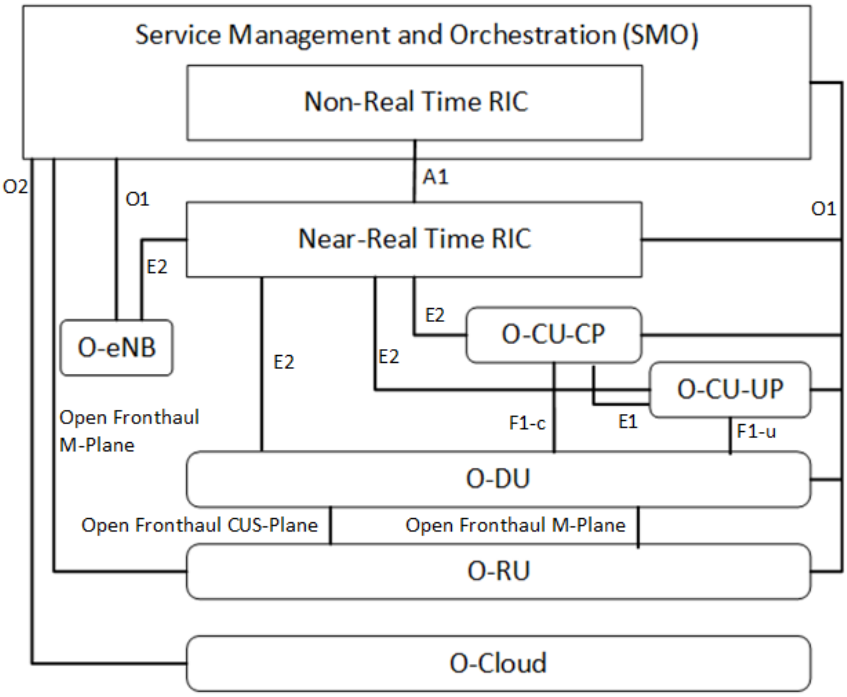

Figure 1.3 Logical architecture of O-RAN [5]

1.1.2 AI in CN domain

For the CN, 3GPP has been working on the development of the Network Data Analytics Function (NWDAF) as a key component of the 5G network architecture since Rel-15 [6]. As shown in Figure 1.4, the NWDAF is responsible for collecting and analyzing network data from CN 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 advantages of the capabilities of 5G networks. In 3GPP, the NWDAF provides Analytics Logical Function (AnLF) and Model Training Logical Function (MTLF) to 5G Core (5GC) network functions (NF), and OAM. Data collection coordination and delivery function coordinate the collection and distribution of data requested by NF consumers.

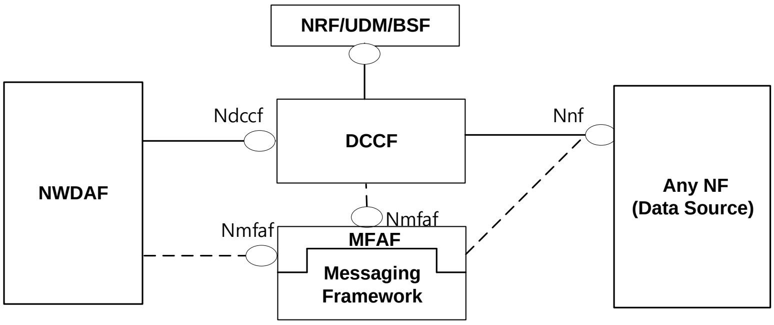

Figure 1.4 Reference architecture for CN data analytics [6]

1.2 Management domain

Currently, AI/ ML is applied in the management domain mainly to solve problems such as intelligent fault diagnosis, resource optimization, and data analysis to help achieve network autonomy. This section describes the relevant research and work done by 3GPP, European Telecommunications Standards Institute (ETSI), and Open Networking Automation Platform (ONAP), in applying AI in the management domain.

1.2.1 3GPP SA5 Management Data Analytics (MDA)

Management Data Analytics (MDA) [7][8] Management Services (MnS) process network and service data as well as events, such as those indicated by performance data (measurements, Key performance indicators (KPIs), Minimization Drive Test (MDT) reports), alarms, and network configuration data. The outputs provided by an MDA Service (MDAS) producer to a service consumer are analytics like predictions, root cause analysis, and action recommendations.

The scope of an MDAS producer can be domain-specific, i.e., either RAN or core network, or cross-domain. Correspondingly, the ingested data and produced analytics results may be related to gNBs and/or specific core network functions.

Figure 1.5 shows cross- and single domain MDAS producer interaction, where the RAN and CN domain MDA MnS produces interface to the respective RAN (gNB) and CN (NWDAF) network functions and their MnS.

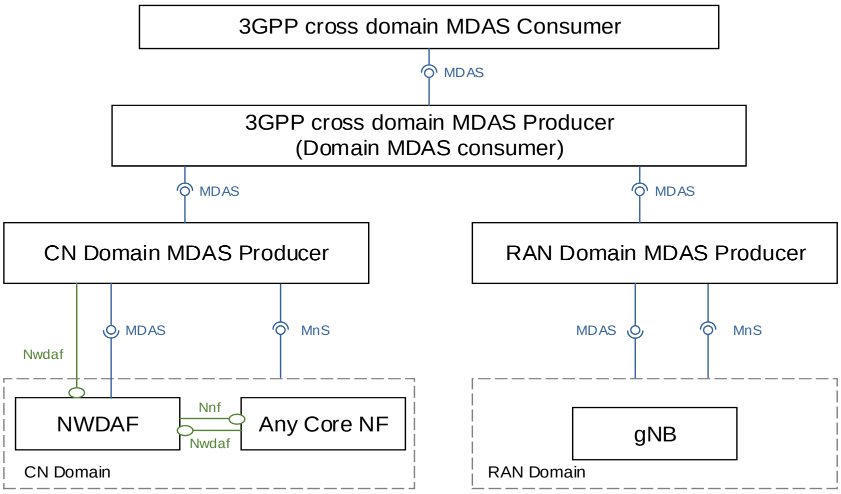

Figure 1.5 Coordination between NWDAF, gNB and MDAS producer [7]

Potential scenarios shown in Figure 1.5 include [7]:

- NWDAF, gNB, and cross-domain MDA consuming MDA MnS provided by the respective domain-specific MDA MnS producer.
- CN domain MDA MnS producer consuming services provided by NWDAF/other CN NFs.
- RAN domain MDA MnS producer consuming MnS provided by the gNB.
- Cross-domain MDA producer consuming MnS/MDA of the RAN and CN domains
- Cross-domain MDA producing MDA MnS for a cross-domain MDA consumer.

1.2.2 3GPP SA5 AI/ML Management

In the study on AI/ML Management [9][10], the generic workflow of the operational steps in the lifecycle of an ML model or entity, is depicted in Figure 1.6. Three use case categories were defined:

- Use cases related to the training phase,
- Use cases related to the inference phase,
- Use cases with requirements in both the training and inference phases.

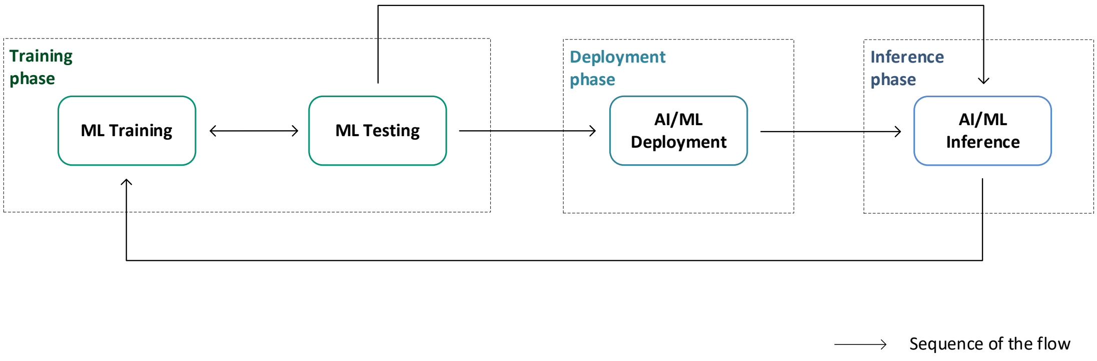

Figure 1.6 AI/ML operational workflow

The use cases focus on ensuring that the functionality containing the AI/ML capabilities has an interface acting as the producer of the AI/ML related MnS.

(1) Management capabilities for training phase

In the training phase, the data used for training the ML entities as well as the processes of training, testing, and validation are managed. ML training data management may include capabilities for processing of data as requested by a training function, by another management function, or by the ML training MnS consumer. Then, based on the availability of training data, ML training management offers capabilities for enabling the MnS consumer to request and manage the model training /re-training. While some basic management capabilities for ML training were already agreed normatively, more capabilities related to ML training have been discussed in the study (critical among them being training configuration and performance management): e.g., the MnS consumer may activate or deactivate the training or configure the ML entity to be trained, the training function or its subsequent processes; the MnS consumer may also manage the training performance and set policies for the training, e.g., for the case that a producer can initiate ML training without explicitly being triggered by the MnS consumer.

Further capabilities deal with the testing and validation of ML entities. ML testing management includes capabilities enabling the MnS consumer to request the ML entity testing, and to receive the testing results for a trained ML model. It may also include capabilities for selecting the specific performance and trustworthiness metrics to be used or reported by the ML testing function on the one hand. On the other hand, ML validation includes the capabilities to evaluate the performance of the ML entity when performing operations on the validation data, and to identify the variability of the performance on the training data and the validation data. For example, if the variability is not acceptable, the entity would need to be tuned (re-trained) before being made available to the MnS consumer and used for inference.

(2) Management capabilities for the inference phase

The main capability in the inference phase is AI/ML inference function control which enables the MnS consumer to control the inference function including the activation and deactivation of the function. This may be extended to include ML entity activation/deactivation by the MnS consumer, including instant activation, partial activation, and schedule-based or policy-based activations. However, part of the AI/ML inference control may include AI/ML deployment control and monitoring which involves capabilities enabling the inference function to load the ML entity. Thereby the MnS producer (inference function) may provide information to the MnS consumer when a new ML entity is available, enabling the MnS consumer to request the loading of the ML entity, the update of the function using the new ML entity or to set the policy for such deployment or update as well as to monitor the corresponding processes.

Another major capability is AI/ML inference orchestration intended to enable the MnS consumer to orchestrate the AI/ML inference functions, given the knowledge of capabilities of the inference functions, the expected and actual running context of ML entity, the AI/ML inference performance, the AI/ML inference trustworthiness, etc. As an example, the MnS consumer may set the conditions to trigger the specific inferences based on the expected outcomes of those inferences.

(3) Management capabilities common to the training and inference phases

The main common capabilities are related to AI/ML trustworthiness management, focused on allowing the MnS consumer to configure, monitor, and evaluate the trustworthiness of an ML entity covering the whole lifecycle. This applies to the ML entity at the training and testing stages as well as when the ML entity is used by an AI/ML inference function.

In addition, common capabilities on configuration management and performance management also apply to both training and inference. AI/ML configuration management is intended to enable the MnS consumer to configure the features of the ML entity or its related training and inference process. AI/ML performance management should enable the MnS consumer to monitor and evaluate the performance of an ML entity at training/testing stages or when used by an AI/ML inference function.

(4) Deployment Scenarios Reference

The ML training or inference function can be located in the cross-domain management system or the domain-specific management system (i.e., a management function for RAN or CN). The ML training function, and AI/ML inference function may be deployed separately from each other, or any two or all of these functions may be co-located. (e.g., Figure 1.7 shows an example of intelligence in RAN where ML training and corresponding management are located at the RAN management function, while AI/ML inference function and corresponding management capability are located locally in gNB.)

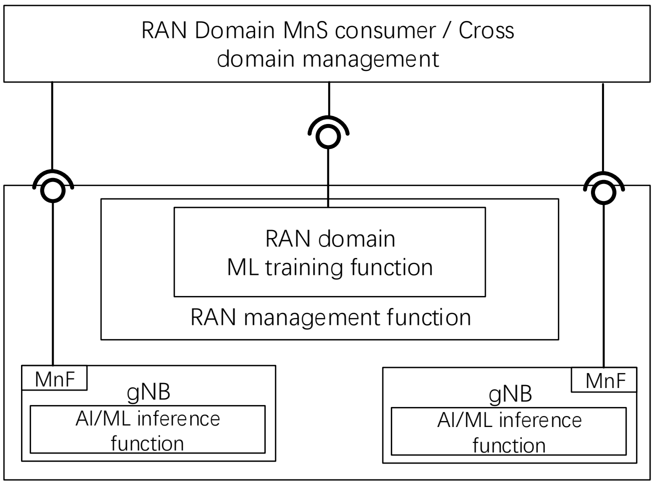

Figure 1.7 RAN domain-specific management function for AI [10]

1.2.3 ETSI ZSM AI Enablers

The goal of ETSI Zero touch network & Service Management (ZSM) is to enable zero-touch automated network and service management in a multi-vendor environment. The ZSM architecture (Figure 1.8) enables exposing and consuming MnS across a set of Management Domains (MDs) and an end-to-end (E2E) Service Management Domain.

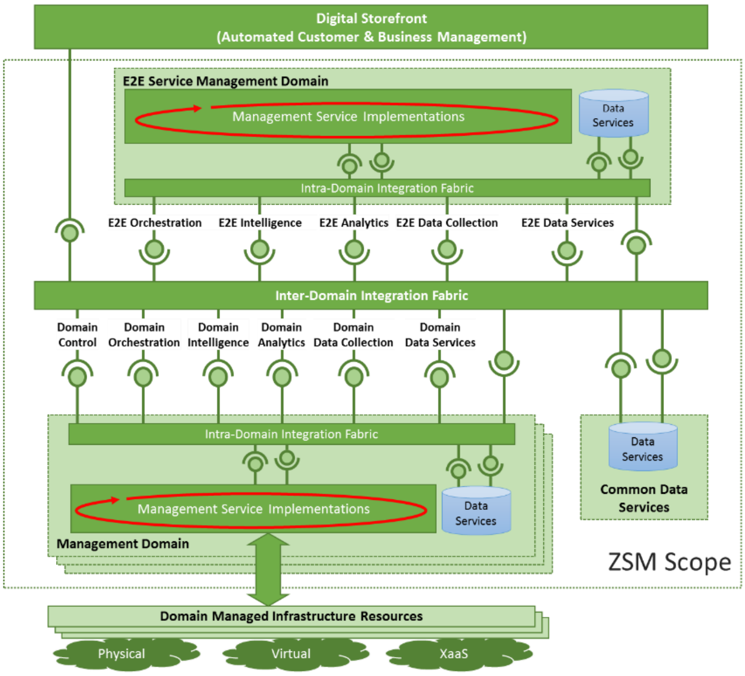

Figure 1.8 ETSI ZSM architectural framework [11]

The technical specification in [11] extends the set of MnS available within the ZSM architecture to address cross-domain AI/ML aspects. In addition to common MnS (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 1.9):

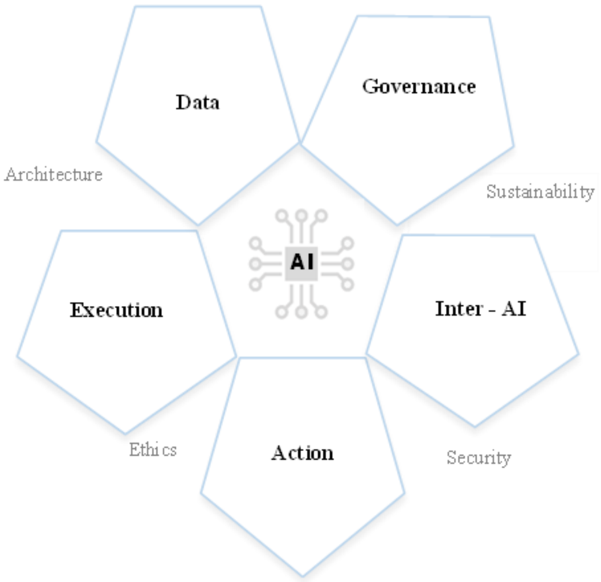

Figure 1.9 ETSI ZSM AI enabling areas [12]

- Data: providing data access across domains and ensuring e.g., data quality, privacy, and security are 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 potentially many types and huge number of AI/ML application instances which is relevant both within a domain and across domains (e.g. MnS configuration of Federated Learning, transfer learning, and other distributed learning.)
- Governance: interfacing the AI/ML-enabled management domains to the human network operator with management services targeting the mapping of Intent to AI/ML applications on the one hand, and instrumenting AI/ML applications to be “trustworthy”, i.e., explainable, robust and fair (MnS Data & Model Trust management, Data & Model Trust evaluation, ML Fallback Management) on the other hand.

Thereby, ZSM enables a range of cross-domain “AI/ML for Network Management” as well as “Management of AI/ML” scenarios like (cf. [12], Annex A) ML-based Anomaly Detection, Federated Learning for Network Management, Trustworthy ML, Distributed ML, ML model validation and ML model cooperation.

1.2.4 ONAP AI Management Practices

ONAP provides a comprehensive platform for real-time, policy-driven service orchestration and automation. The Data Collection, Analytics, and Events (DCAE) project provides intelligence for ONAP to support automation by performing network data collections and analytics, which includes Micro-Services (MS), i.e., collectors, analytics, and event processors to support active dataflow and processing. Among these MSs, Slice Analysis MS is a key component, which is used to perform E2E intelligent slicing. The functions can be summarized as follows：

- Analyzing the Fault Management (FM) / Performance Management (PM) data and KPI data related to various slice instances, slice sub-net instances, and services catered by the slices.
- Determining and triggering appropriate control-loop actions based on the analysis above.
- Receiving recommendations for closed-loop actions from AI/ML or Analytics engines.

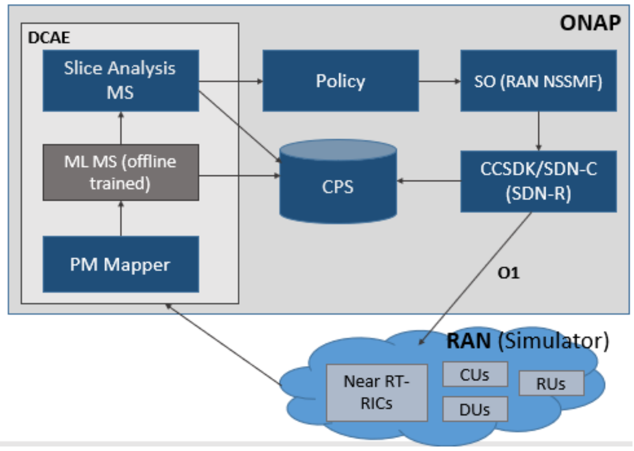

Figure 1.10 ML MS enhancement on intelligent slicing [13]

Error! Reference source not found. provides an example of ML MS enhancement on intelligent slicing, where Slice Analysis MS can consume data including a list of Near-RT RICs and the current configuration of the Near-RT RICs. Based on the data, Slice Analysis MS computes the slice performance-related value, and the computed value is compared with the current configuration of the Near-RT RICs. If the change in configuration exceeds the minimum percentage value, which is kept as a configuration parameter, the closed-loop will be triggered. Upon reception of the recommendation to update the configuration of RAN from AI/ML or Analytics engines, the Slice Analysis MS prepares and sends a control loop onset message.

2 Use cases and considerations of cross-domain AI

As network application types and scenarios become more complex, future networks need to provide ubiquitous native AI capabilities. This chapter identifies some use cases recommended being enabled by AI collaboration across different domains. Then, we further give some technical considerations and requirements for cross-domain AI.

2.1 Use cases

2.1.1 Cross-domain data analysis

With the continuous enrichment of application types, various domains in future networks will generate a large amount of data. Cross-domain data analysis services can utilize AI/ML to analyze ubiquitous data and provide end-to-end analysis reports and optimization suggestions to ensure differentiated service requirements for users. Currently, 3GPP MDAS can collect and analyze management-related data from RAN, CN, and other domains to provide data analysis services in various scenarios. For example, it can offer services for mobility management, energy efficiency analysis, Service Level Specification (SLS) analysis, and more. Table 2.1 shows an example of inputs and outputs related to AI in the RAN, CN, and management system for mobility management scenarios. Cross-domain data analysis services collect measurement information from the RAN and CN, and provide analysis reports and recommendations to the gNB or other NFs. Using AI technology to analyze data generated at various locations in the network can help optimize the operational efficiency and enhance the user experience.

Table 2.1 Cross-domain data analysis for mobility management

	Data Source	**Input**	**Output**
MDAS[7]	UE, NWDAF, gNB, etc.	1. Performance Measurements (Reference Signal Received Power (RSRP) / Reference Signal Received Quality (RSRQ) / Signal to Inference plus Noise Ratio (SINR), End-to-end Latency, Throughput, Data packet loss…) 2. Resource information 3. UE location report…	1. Consumed/projected resource information. 2. Priority of the target gNB for optimal handover (HO). 3. Recommendation for gNB modification
NWDAF[6]	Access and Mobility Management Function (AMF)	1. UE locations (UE location, Timestamp) 2. Type allocation code(TAC) 3. UE access behavior trends 4. UE location trends	UE mobility statistics or predictions 1. Time slot start, duration 2. UE location: Observed location statistics/ predicted locations 3. Confidence in prediction
RAN[2]	UE	1. UE location information 2. Radio measurements (RSRPs/ RSRQs/ SINRs)	1. UE trajectory prediction 2. Estimated arrived probability in (Conditional handover) CHO 3. Predicted handover target node, candidate cells in CHO 4. UE traffic prediction
	Neighboring RAN Node	1. Position, QoS parameters of historical HO-ed UE 2. Current/predicted resource status…	
	Local Node	1. UE trajectory prediction 2. Current/predicted resource status and traffic	

Based on the O-RAN architecture, Near-RT RIC can serve as the agent for RAN domain data analysis, and SMO needs to add a cross-domain data analysis producer to collaborate with the core network NWDAF and Non-RT RIC to achieve cross-domain intelligent data analysis. Future considerations include:

1. How to subscribe to data from different network domains? What interfaces/methods can be used?

2. Data types, dimensions, and structures of the data collected from different domains may vary. How to preprocess, clean, and integrate the data?

3. How to reduce communication overhead during cross-domain data transfer and ensure data security and privacy?

4. How to evaluate the performance of the cross-domain data analysis model. How to provide feedback and update the models?

5. How to ensure data alignment when using distributed learning?

2.1.2 Intent based E2E smart slicing (ONAP)

For scenarios requiring a new E2E slice [14], network slicing management system slices the SLA (such as network slicing list, PLMN list, maximum number of users, service areas, the end-to-end delay) for each domain (access network, transmission network, core network, etc.) of the SLA decomposition according to the requirements of the tenants. Then each domain can utilize AI/ML for resource configuration, including bandwidth, delay and so on. In the process of slicing operation, operators are unable to effectively evaluate the service quality of various services due to the lack of real-time monitoring service related experience. Thus, an intelligent network slicing method is urgently needed. By evaluating the SLA of the network slicing, the service experience in the slicing can be monitored in real time, so as to accurately perceive the business quality experienced by users and make accurate, dynamic adjustments and error corrections of the network.

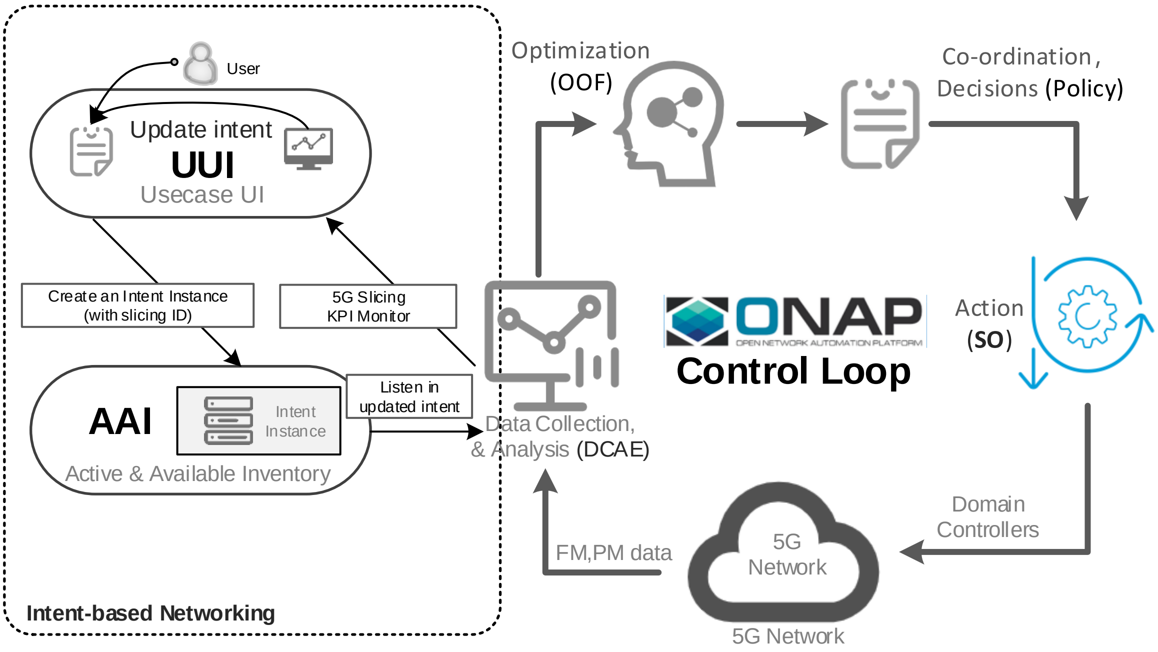

Figure 2.1 A closed-loop structure for E2E slicing [15]

ONAP has provided an example of smart E2E slicing which is given in Figure 2.1. The ONAP Operation Manager (OOM) is a microservice deployment system based on Kubernetes. Table 2.2 lists the microservices which are related to the E2E slicing use cases.

Table 2.2 ONAP microservices which is related to the E2E slicing use cases

Short Name	Full Name	Description
UUI	Use case User Interface	Provides the user interface for users to interact with ONAP
SO	Service Orchestrator	Provides end-to-end service orchestration
DCAE	Data Collection, Analytics, and Events	Performs data analytics of the telemetry data
Policy	Policy Framework	Executes policies
AAI	Active and Available Inventory	Data store for network and service configurations, network resources, inventory, etc.

The network sends FM and PM data to the DCAE subsystem to monitor the network state. When the network state cannot meet the user's demand, DCAE can modify the network policy configuration to achieve the modification of the slice instance or policy to achieve the modification of the slice parameters. The modification request is executed through SO, and then the network is controlled in the domain controller (RAN domain, CN domain, TN domain) to make some modifications to the network state.

Therefore, in order to achieve intent-driven, end-to-end, intelligent slicing, the management system needs capabilities such as RAN and CN data collection, intent transfer and decomposition, and cross-domain AI orchestration. By leveraging cross-domain AI collaboration, end-to-end slicing requirements can be guaranteed, while also enabling real-time monitoring and flexible adjustment of slicing performance.

2.1.3 Network Digital Twins

NDT in the realm of Open RAN heralds a new era of cross-domain network management and optimization. These virtual replicas encompass various network domains (access, core, transport, and application) providing a holistic view of the network's performance and interactions. In multi-domain environments, NDT offers enhanced capabilities for addressing use cases in optimization, operation & management, and planning. In essence, NDT in open RAN presents a powerful multi-domain approach to modernize network management and optimization. Multi-domain network digital twin in Next Generation (Next G) networks can be applied to various use-cases, encompassing optimization, operation, and testing, as well as planning. Here are some examples:

(1) Optimization Use-Cases:

- **Resource Allocation Optimization:** The network digital twin can optimize resource allocation across multiple domains, such as compute, communication, and storage, to maximize network efficiency and performance. It can dynamically allocate resources based on real-time demand, traffic patterns, and QoS requirements, ensuring optimal resource utilization.

  The network digital twin plays a crucial role in facilitating cross-domain resource allocation optimization in next-generation networks. This is how the network digital twin contributes to this domain:

  Holistic Resource View: The network digital twin provides a comprehensive and real-time view of the network infrastructure and its available resources across multiple domains. It includes information about compute resources, communication elements, storage capacities, and other network components. This holistic view allows network operators to have a clear understanding of the available resources and their utilization across different domains.

  Resource Modeling and Simulation: The digital twin enables the modeling and simulation of different resource allocation scenarios. It can simulate resource demands, traffic patterns, and application requirements to assess the impact of various allocation strategies. By analyzing these simulations, operators can identify optimal resource allocation configurations that maximize network efficiency, performance, and resource utilization.

  Dynamic Resource Optimization: With the NDT, operators can dynamically optimize resource allocation based on real-time network conditions and demands.

By continuously monitoring the actual resource utilization and comparing it with the digital twin model, operators can identify areas of overutilization or underutilization. This information allows them to make real-time adjustments to resource allocations, ensuring efficient utilization of resources across domains.

Quality of Service Optimization: The digital twin assists in optimizing resource allocation to meet the QoS requirements of different applications and services. By simulating and analyzing the impact of resource allocation on performance metrics such as latency, throughput, and reliability, operators can identify resource allocation strategies that prioritize critical services while balancing the needs of other applications. This ensures that the network delivers the required QoS levels across domains.

Load Balancing and Traffic Optimization: The network digital twin helps in load balancing and traffic optimization by providing insights into network traffic patterns and resource utilization. By analyzing these patterns, operators can identify congested areas or underutilized resources and make adjustments to balance the load across domains. This optimization technique improves overall network performance and ensures efficient resource utilization.

Real-Time Resource Monitoring and Management: The digital twin enables real-time monitoring of resource usage and performance metrics across different domains. It provides operators with up-to-date information on resource availability, usage patterns, and performance bottlenecks. This real-time visibility allows them to identify potential resource allocation issues and take proactive measures to address them promptly.

Optimization Across Heterogeneous Networks: Next-generation networks are expected to be heterogeneous, comprising different types of networks such as cellular, Wi-Fi, and satellite.

The NDT helps optimizing resource allocation across these heterogeneous networks by considering their unique characteristics and capabilities. It enables operators to leverage the strengths of each network type and allocate resources accordingly, ensuring efficient utilization and improved overall network performance.

These capabilities empower network operators to allocate resources efficiently, optimize network performance, and meet the diverse requirements of applications and services in next-generation networks.

- **Energy Efficiency Optimization:** The NDT serves as a powerful tool in simulating and analyzing energy consumption across diverse domains, pinpointing avenues for energy refinement. Beyond energy efficiency, NDT-enabled infrastructure is designed for vital operational, testing, and optimization use-cases. It supports advanced strategies like dynamic power management, load balancing, and energy-efficient routing to diminish energy use and lower operational costs. While it is acknowledged that the NDT itself requires a substantial energy investment, the trade-off becomes evident when considering its broader applications. The energy consumed by the NDT is offset by the improvements and efficiencies realized across the three core use-case categories: operation and testing, planning and optimization. Thus, the comprehensive advantages offered by the NDT, spanning from enhanced energy efficiency to robust optimization capabilities, often justify the energy costs associated with its operation. By producing a virtual mirror of the network infrastructure—encompassing compute resources, communication elements, and power systems—the NDT facilitates real-time monitoring, evaluation, and optimization of energy consumption across various domains. Here is how it helps in improving energy efficiency:

Energy Monitoring and Analytics: NDT continuously collects data on energy consumption from various network components. This includes information on power usage, resource utilization, traffic patterns, and environmental conditions. By analyzing this data, energy hotspots, inefficiencies, and potential areas for energy optimization can be identified. Insights gained from the NDT can guide decision-making processes for energy-efficient resource allocation and operation.

Energy-Aware Resource Allocation: The NDT can provide guidance on resource allocation decisions by considering energy efficiency as a key objective. By simulating different scenarios and assessing their energy implications, the digital twin can recommend resource allocation strategies that minimize energy consumption while meeting performance requirements. For example, it can suggest intelligent placement of compute tasks, adjustment of communication parameters, or load balancing techniques to optimize energy efficiency.

Dynamic Energy Management: The NDT allows for real-time energy management based on dynamic network conditions. By integrating real-time data from sensors and devices, the digital twin can adapt energy consumption based on the current network demands. For instance, during periods of low traffic or idle compute resources, the digital twin can recommend power-saving modes, such as sleep or low-power states, to conserve energy while maintaining network readiness.

Energy Optimization Strategies: Leveraging the insights provided by the network digital twin, energy optimization strategies can be developed and implemented across different domains. These strategies may include dynamic power control, intelligent resource scheduling, workload consolidation, and traffic optimization techniques. The digital twin facilitates the evaluation and fine-tuning of these strategies, ensuring their effectiveness in improving energy efficiency across the network.

Predictive Maintenance and Energy Planning: By simulating various scenarios and predicting future energy requirements, the network digital twin assists in proactive energy planning and maintenance. It can identify potential energy-related issues, such as overutilization, inefficient power distribution, or cooling inefficiencies. By addressing these issues in advance, network operators can optimize energy usage, reduce downtime, and improve overall energy efficiency.

This results in reduced energy costs, minimized environmental impact, and enhanced sustainability of Next G networks.

- **Service Orchestration and Network Slicing:** By modeling and simulating network slices, the digital twin can facilitate efficient service orchestration. It can optimize the allocation of resources, ensure isolation between slices, and dynamically adjust resource allocations based on changing service requirements. NDT assisted Service Orchestration and Network Slicing could be performed as follows:

  Modeling and Simulation: The network digital twin provides a virtual representation of the network infrastructure, including compute, communication, and storage resources across different domains. It allows for modeling and simulating various network slices and service scenarios. This enables network operators to evaluate the feasibility and performance of different service orchestration and network slicing configurations before actual deployment.

  Resource Allocation and Optimization: With the network digital twin, operators can optimize the allocation of resources across multiple domains to support network slices efficiently. By simulating resource demands, traffic patterns, and QoS requirements, the digital twin assists in determining the appropriate allocation of compute, communication, and storage resources to meet the needs of each network slice. It helps ensure that the resources are efficiently utilized, and the QoS requirements of different services are satisfied.

  Dynamic Slice Management: The digital twin enables dynamic management of network slices by continuously monitoring their performance and adjusting resource allocations as needed. It provides real-time insights into the utilization and performance of each network slice, allowing operators to make informed decisions on scaling resources, adjusting network parameters, and ensuring optimal slice operation. This dynamic management capability enables efficient service orchestration and adaptation to change service demands.

  Service Isolation and Security: Network slicing requires strong isolation between different services to ensure data privacy, security, and QoS guarantees. The digital twin assists in modeling and enforcing isolation mechanisms to prevent interference or unauthorized access between network slices. It helps in identifying potential security vulnerabilities, simulating threat scenarios, and optimizing security measures to protect the integrity and confidentiality of each network slice.

  Service Level Agreement (SLA) Management: The network digital twin supports SLA management by providing a holistic view of the network slices and their performance metrics. It helps in monitoring and enforcing SLAs, ensuring that the agreed-upon service guarantees are met. The digital twin enables operators to track KPIs of each network slice, identify deviations from SLAs, and take proactive measures to maintain service quality.

  Fault Detection and Self-healing: The digital twin can monitor the health and performance of network slices in real-time. By comparing the actual state of the network with the digital twin model, it can detect anomalies, failures, or performance degradation. This information enables operators to trigger self-healing mechanisms, such as automatic reconfiguration, resource reallocation, or failover mechanisms, to ensure uninterrupted service and maintain the desired service levels.

The network digital twin empowers efficient cross-domain service orchestration and network slicing by providing modeling and simulation capabilities, optimizing resource allocation, enabling dynamic slice management, ensuring service isolation and security, supporting SLA management, and facilitating fault detection and self-healing. It serves as a powerful tool for network operators to design, deploy, and manage network slices effectively, meeting the diverse service requirements and enabling the flexible and scalable delivery of services in next-generation networks.

(2) Operation and Testing Use-Cases:

- **Fault Detection and Self-healing:** The network digital twin can monitor network performance and detect anomalies or failures in real-time. By comparing the real network state with the digital twin model, it can identify issues and trigger proactive self-healing mechanisms to maintain uninterrupted service and minimize downtime.
- **Network Performance Testing:** The digital twin can be used to simulate and test different network scenarios before actual deployment or major changes. It allows network operators to assess the impact of new services, applications, or network configurations, optimizing performance and ensuring smooth operation.
- **Security and Threat Analysis:** The digital twin can simulate potential security threats and vulnerabilities, allowing for proactive detection and prevention measures. It can assist in evaluating the effectiveness of security mechanisms, identifying potential weaknesses, and improving network resilience against cyber threats.

(3) Planning Use-Cases:

- **Network Capacity Planning:** The network digital twin can assist in planning network capacity to accommodate increasing traffic demands. By modeling different network elements and simulating traffic patterns, it helps determine optimal capacity requirements, identify potential bottlenecks, and plan for network expansion or upgrades.
- **Coverage and Deployment Planning:** The digital twin can simulate coverage maps and assess the impact of different deployment scenarios. It assists in determining the optimal placement of base stations, access points, and other network elements to achieve desired coverage, capacity, and Quality of Experience (QoE) metrics.
- **Spectrum Management and Optimization:** The digital twin can analyze spectrum availability and predict interference scenarios. It aids in optimizing spectrum allocation, evaluating the impact of new frequency bands, and improving spectrum efficiency for better network performance.

By leveraging multi-domain network digital twin technology, operators can optimize their networks, enhance operational efficiency, and make informed decisions in areas such as resource allocation, energy management, fault detection, performance testing, security, and network planning. These use-cases enable more efficient, resilient, and adaptive 6G/5G networks capable of meeting the evolving demands of diverse applications and services.

2.1.4 Cross-domain QoE estimation

Estimating end-user perceived quality from an application service has been the focus of mobile network operators, as end-user customer satisfaction impacts network operator revenue.

Well-known QoS is a function of QoE, where QoE considers performance estimation at a higher layer in the network stack (closer to the user). In order to model and assess QoE accurately, it is important to include observation attributes (metrics indicating so-called QoE factors) from different measurement points in the pipeline. For the assessment of video QoE during a video streaming application running at the end-user device (UE), video playout bitrate, stalling event duration and frequency, user profile, user experience, and video codec can be important metrics; from radio access network, a received signal strength quality such as RSRP, RSRQ, SINR, interference, cell load are the important ones; while in the application server, some metrics related to QoE could be the application server load, inter-departure time of the video packets, etc.

These datasets are inherently decentralized, as they are naturally collected and originated from different physical data sources. In conventional centralized ML, all observations obtained over distributed measurement points at different network layers (radio, network, application) are delivered over a communication channel to one centralized location that has good hardware resources. Then, an ML model is trained. However, there are scenarios where moving those decentralized datasets to a central location is difficult (if not impossible). Some reasons may include data privacy, high data volume, prohibitive delay, and unavailable bandwidth for data transfer. Thanks to distributed ML approaches, a large global model can be split and placed partially on distributed nodes enabling decentralized cross-domain learning. For example, for QoE estimation across terminals, RAN, CN, and application domains, a Vertical Federated Learning approach can be used. At this point, model transfer and computing resource scheduling across different domains need to be considered. Therefore, intelligent collaboration across domains is an important enabling technology for QoE estimation.

2.2 Technical considerations of cross-domain AI

We have explored different use cases and hinted at enabling technologies driving cross-domain AI (e.g., federated learning). In general, when looking at multi-vendor aspects impacting the standard, research should not be limited to looking at how the technology will work. Other important aspects to consider from the beginning are:

1. Intellectual property (IP): in multi-vendor cross-domain AI use cases, how is IP guaranteed?

2. Liability: If something does not work well in a part of the cross-domain solution (e.g., bad performance of a part of a model), who is responsible for what?

3. Innovation: Once an interface has been standardized, it is difficult to change. How to promote cross-domain AI solutions without hindering innovation?

4. Data security and safety: When data is exchanged across different domains, what mechanisms can be used to ensure data security and privacy?

5. Exposure capability: Cross-domain data sharing and cross-domain AI internal capabilities exposure can be provided.

6. Cross-domain management capabilities: Cross-domain management and orchestration capabilities can be provided in order to effectively manage and maintain the life cycle AI/ML models.

7. Computing resources: cross-domain AI use cases can take into account the availability of resource capability of network nodes such as Central Processing Unit (CPU)/Graphics Processing Unit (GPU). This is especially true for distributed AI operations where involved nodes are e.g., UEs, Central Units (CUs)/Distributed Units (Dus).

8. Data: Cross-domain AI use cases require a distributed data-driven architecture that is able to connect distributed data domains. Data can be secured via access permission and data management can follow regulations for data protection if required.

3 Potential Solutions for Cross-Domain AI

3.1 Impact on network architecture.

We have investigated potential technology directions to enable cross-domain AI in order to meet the aforementioned requirements. Figure 3.1 provides two possible solutions that are built on the O-RAN architecture. The first one focuses on edge network scenarios, where a RAN-CN converged architecture can naturally enhance collaboration between RAN and CN domain AI while minimizing the overhead associated with information flow via the X2 interface. The second solution involves leveraging the management domain to enable the collaboration and management of multi-domain AI. This can be achieved by introducing a cross-domain AI controller in the SMO, which interacts with Non-RT RIC, core network, and other domain-specific AI controllers.

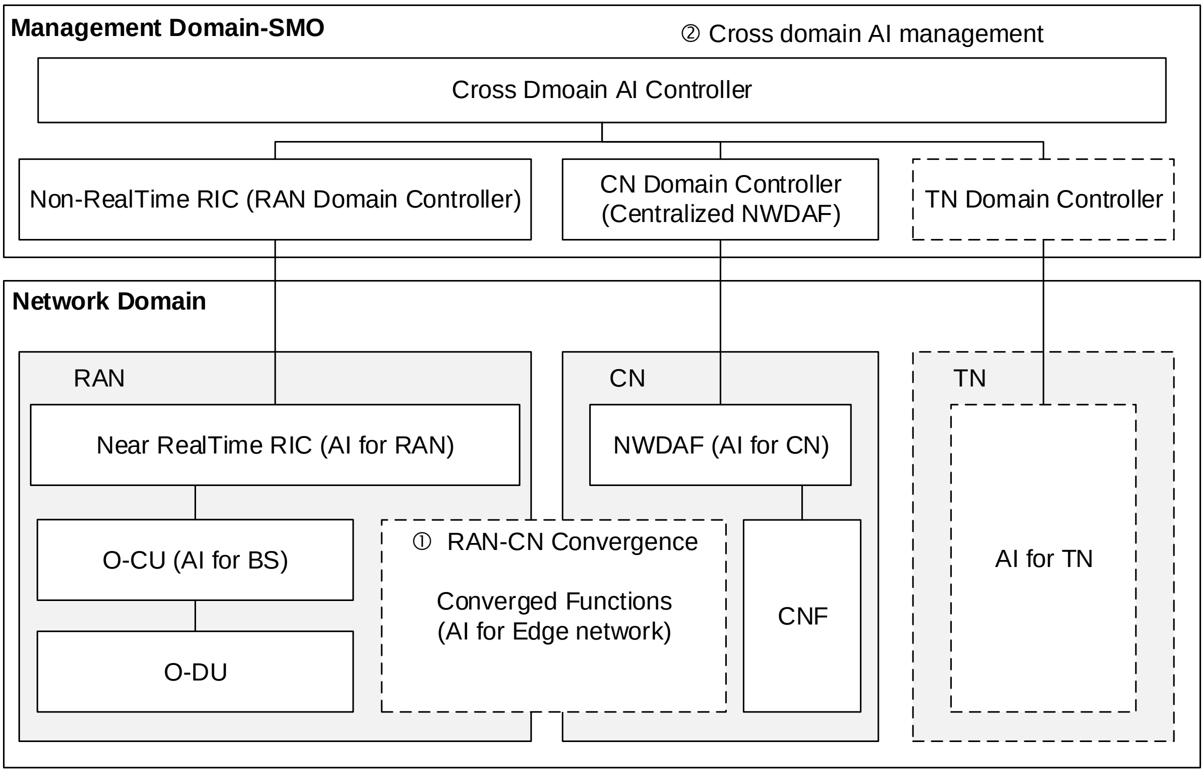
Management Domain-SMO
② Cross domain AI management
Cross Dmoain AI Controller
Non-RealTime RIC (RAN Domain Controller)
CN Domain Controller (Centralized NWDAF)
TN Domain Controller
Network Domain
RAN
Near RealTime RIC (AI for RAN)
O-CU (AI for BS)
O-DU
① RAN-CN Convergence
Converged Functions (AI for Edge network)
CN
NWDAF (AI for CN)
CNF
TN
AI for TN

Figure 3.1 Potential solutions and architecture for cross-domain AI

3.1.1 Converged RAN-CN architecture

In wireless networks, the collaborative efforts of CN and RAN are essential to facilitate a wide range of AI use cases with distinct requirements. For instance, the optimization of mobility management parameters, radio resource management parameters, slice resource allocation, and end-to-end QoS assurance necessitates the collection of data from both CN and RAN, followed by the implementation of separate policies. However, the existing architecture lacks adequate provisions for cross-domain AI. Firstly, the existing architecture of RAN and CN poses obstacles to the exchange of cross-domain data information. Specifically, RAN can only connect to the AMF in the CN control plane through a point-to-point N2 interface, which necessitates the AMF's involvement in the forwarding of control data between RAN and NWDAF, resulting in increased transmission delays and overhead. Moreover, there is a lack of standardized interfaces to facilitate Near-RT RIC/Non-RT RIC interaction with NWDAF. Secondly, considering the distributed intelligent deployment, RAN and CN's intelligent functions are physically proximate, presenting an opportunity for further integration of logical functions between the two. This potential fusion can simplify the network architecture and interfaces, eliminate functional redundancy, reduce latency, and mitigate unnecessary forwarding overhead. To achieve RAN-CN collaboration/convergence for endogenous AI, the following aspects need to be considered from the perspective of O-RAN:

- Consider the analysis of collaboration/convergence requirements and use cases, as well as the identification of specific collaboration/convergence functions to facilitate the sharing and coordination of data, models, algorithms, etc. between RAN and CN.

- Identify the impact of collaboration/convergence on O-RAN architecture, protocols, interfaces, and procedures. More flexible and efficient O-RAN and RAN architectures, such as Service-based Architecture (SA), can be explored to better accommodate and support RAN-CN collaboration/convergence. Appropriate protocols and interfaces should be devised to facilitate data transmission and control. In light of this, the data transmission and collaboration workflows subsequent to collaboration/convergence should also be taken into account.
- Explore data security and privacy protection considerations. RAN-CN collaboration/convergence entails extensive data sharing and processing, making the security and privacy of the data paramount. It is crucial to employ suitable security mechanisms and privacy protection strategies to safeguard the confidentiality and integrity of user data and network information.

3.1.2 Enhanced SMO for cross-domain AI

Currently, SMO is capable of providing AI training and model management functions, and it sends policy to Near-RT RIC through the A1 interface. However, in order to enable coordinated AI capabilities across different network domains, the following functionalities could be added to the current SMO:

- A cross-domain AI control function to handle the management and orchestration of AI across network domains.
- Capability to obtain AI capability information from different network domains and enable AI capability exposure.
- End-to-end AI management and orchestration capability, such as cross-domain data arrangement and mapping, AI task identification and decomposition, deployment to matching computing nodes, etc.
- In addition to data collection, model training, and inference, the management of model generation, storage, and migration, as well as processes such as AI performance evaluation and model updates, are necessary to enhance the end-to-end lifecycle management capabilities of AI.
- Support for cross-domain distributed learning, for example, by managing the training and inference of vertical federated learning.
- Capabilities of intent management, such as intent recognition, translation, and matching ability, intent conflict resolution capability, and closed-loop control capability in a cross-domain AI context.
- Support energy saving and energy efficiency through cross-domain AI.

3.2 AI elements collaboration across different network domains

3.2.1 Data

Data is one of the basic elements of AI technology, and achieving cross-domain AI capabilities requires the collection and processing of data across different domains. The vast amount of data in future networks serves as the foundational basis for cross-domain AI, while advancements in data transmission and processing technologies provide the necessary technical infrastructure for the application of cross-domain data. However, it is important to recognize that significant challenges exist with regard to ensuring data security and privacy during cross-domain data collection and usage.

The process of cross-domain AI data collection typically involves decomposing the data requirements of cross-domain AI according to the domain of the data source, generating cross-domain data collection requirements, distributing them to the target domain, and then the target domain collects data in accordance with the domain's specific collection requirements. Finally, the entire process of the target domain exposes the collected data to other domains with consideration of data security and privacy. Specifically, the cross-domain data sharing procedure must at least include the authentication and verification process for the demander.

The cross-domain data processing usually refers to the whole process of handling the data collected during the cross-domain data collection process. It generally includes data pre-processing processes such as data cleansing, data integration, data conversion, data reduction, etc. and data storage process, data using process, and finally data disposal process. This process is mainly based on the characteristics of mobile communication network data and integrates big data technology and communication technology. The Figure 3.2 below is a typical example of a cross-domain data processing process.

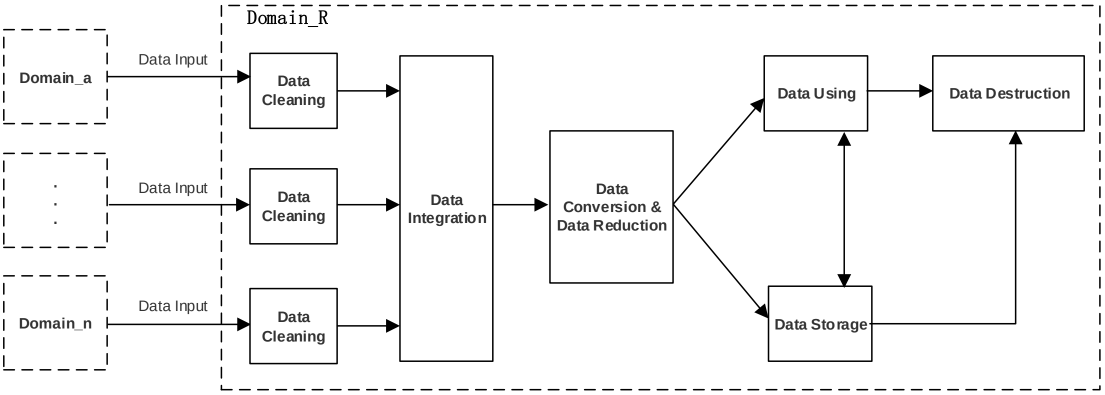

Figure 3.2 An example of a cross-domain data processing process

This example includes the following data processing methods:

- Data Cleaning: the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- Data Integration: the process of combining data from different sources to eliminate duplicate and inconsistent data.

- Data Conversion: the process of translating data from one format to another through normalization, standardization, discretization, and other conversion operations to better support data analysis and modeling.
- Data Reduction: the process of processing data through compression, sampling, feature selection, and other conversion operations to better support data analysis and modeling.
- Data Storage: the process of storing the data on storage media for future use.
- Data Destruction: the process of destroying useless, redundant or old data to prevent nefarious data exploitation and reduce the security risks.
- Data Using: the process of using the data required by cross-domain AI to achieve cross-domain AI capabilities.

This is a typical cross-domain data processing process. In practical applications, data preprocessing methods may be different according to the characteristics of data collected. For example, in the O-RAN architecture, measurement data in CU/DU, etc. (network domain) can be collected by SMO (management domain) with different methods (Refer to section 2.2.2 in [16], Therefore, the data preprocessing operations for data collected by different methods will be different before they are delivered to Non-RT RIC for use.

3.2.2 Algorithm/model

(1) Common AI algorithms and applications:

Due to AI/ML's promising potential in optimizing planning and operations of mobile communication networks, 3GPP initiated "*Study on enhancement for Data Collection for NR and EN-DC*" [3] in Rel-17 and both – "*Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface*" **Error! Reference source not found.** and "Study on Artificial Intelligence/Machine Learning (AI/ ML) management" [10] in Rel-18. O-RAN also treats AI/ML as a strategic topic and multiple O-RAN working groups, e.g., WG2 (Non-RT RIC) and WG3 (Near-RT RIC), and next generation research group (nGRG) are active in introducing AI/ML features to O-RAN.

Efforts in 3GPP and O-RAN focus on facilitating data collection and analytics [6] [7] [8], defining AI/ML terminology including ML model [5] [9] and ML model management (training, deployment, inference, performance evaluation/feedback and update over standard interfaces) [4] [9]. ML models and algorithms are usually treated as a blackbox. They are kept out of the scope of standardization to allow rapid innovation and vendor differentiation. However, ML models and algorithms are a critical part of the entire AI/ML value chain, so this section attempts to describe them briefly.

ML models and algorithms usually leverage one or more of the following learning methods:

- Supervised Learning including Semi-Supervised Learning – learning from labeled training data, where input-output pairs are provided.
- Unsupervised Learning – learning patterns from unlabeled data, discovering inherent structures or relationships.
- Reinforcement Learning – learning based on identification of rewards for a set of actions and system states during exploration phase and maximizing reward during exploitation phase. This can include supervised learning and deep learning algorithms,
- Deep Learning – learning using multi-layer neural networks like the human brain. It is a flavor of supervised learning.

Following learning techniques could leverage ML models and algorithms:

- Transfer Learning – imparting learning from one context to another similar context.
- Federated Learning – learning across multiple nodes keeping data local thus privacy intact at each node. It also saves on network bandwidth and could be useful in cross-domain AI (e.g., across (i) UE and RAN, (ii) RAN and CN, (iii) UE, RAN, and CN and finally, (iv) UE, RAN, CN, and management). Currently, 3GPP is considering a one-sided model (e.g., on UE side, or RAN side, or both, with no interaction) and a two-sided model (e.g., on UE side and RAN side models interacting with each other).

Based on the application, ML models and algorithms could be used for:

- Prediction e.g., Time Series Analysis like Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) networks, Large Language Model (LLM) based transformers.
- Classification (or clustering) e.g., K-means clustering.
- Natural Language Processing (NLP) e.g., Language Translation, Named Entity Recognition
- Computer Vision e.g., Object Detection, Facial Recognition
- Anomaly Detection e.g., Autoencoders
- Recommendation Systems e.g., Content-Based Filtering

Prediction is widely used in mobile communication network context. Some examples of prediction include Cell Based UE Trajectory Prediction, Predicted Radio Resource Status, Predicted Number of Active UEs, Predicted RRC connections, and Energy Cost as proposed in [17].

With the rising popularity of Chat Generative Pre-Trained Transformer (ChatGPT) like AI/ML based applications, Generative Models are becoming increasingly relevant. Some examples include Transformers, Generative Adversarial Networks (GANs).

Depending on the use case, one or more appropriate ML models and algorithms are implemented.

ML models are usually deployed as an image, an executable, a set of source files including metadata, or any other suitable means.

The following are potential research topics for future models and algorithms:

- Application of recent innovations like LLMs (used by ChatGPT)/GANs, Reinforcement Learning (RL), Transfer Learning (TL) and Knowledge Distillation (KD) in mobile communication network.
- Trustworthy AI/ML including Explainable AI/ML. While internals of ML models and algorithms would continue to remain proprietary, it is paramount to clearly explain input-output relationship. Explainable AI/ML plays a critical role in this context. Trustworthy AI/ML preserves user and data privacy as well as the secrecy of ML models and algorithms. It is evident that Trustworthy AI/ML is critical to build trust and confidence about usage of AI/ML in mobile communication networks among telecom operators, regulators, and even end consumers.
- More energy efficient ML models and algorithms to meet the future energy saving and energy efficiency goals.
- Algorithms and models that help realize Native AI.
- Further accuracy improvement of prediction models and algorithms.
- With the advancement in ML models and algorithms, complexity needs to be kept to a manageable level. In addition, the scalability of these models and algorithms needs to be ensured for wider adoption.
- New techniques such as continual learning, causal inference, inverse RL.

Future ML models and algorithms would need to integrate all these novel and necessary concepts and aspects in their requirement, design, and implementation phases.

(2) Model distribution and sharing:

3GPP SA1 has studied the general principle of AI/ML model/data distribution and sharing over 5G system in TR 22.874 [18]. An example of split AI/ML inference across different domains can be depicted in Figure 3.3.

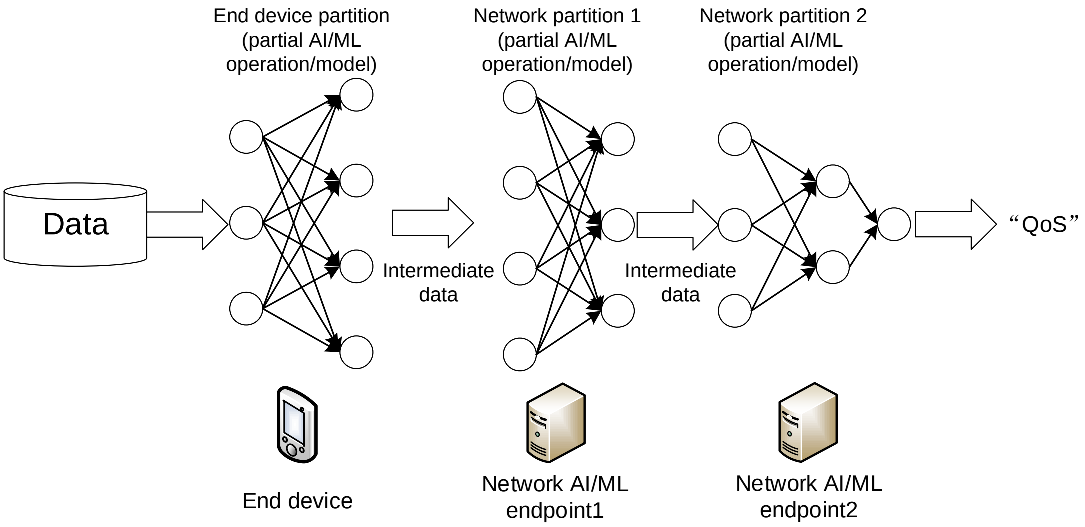

Figure 3.3 An example of split AI/ML inference

The following modes attempt to split the AI/ML inference or even the model into multiple parts (e.g. cross domains) according to the current task and environment, to alleviate the pressure of computation, memory/storage, power, and required data rate on both device and network endpoints, as well as to obtain a better model inference performance on latency, accuracy and privacy protection.

•Mode a): Device-cloud/edge split inference

In this mode, an AI/ML inference operation or model is firstly split into two parts between the device and the cloud/edge server according to the current system environmental factors such as communications data rate, device resource, and server workload. Then, the device will execute the AI/ML inference up to a specific part or the Deep Neural Network (DNN) model up to a specific layer and send the intermediate data to the cloud/edge server. The server will execute the remaining parts/layers and send the inference results to the device.

•Mode b): Edge-cloud split inference

In this mode, the DNN model is executed through edge-cloud synergy, rather than executed only on either cloud or edge server. The latency-sensitive part of an AI/ML inference operation or layers of an AI/ML model can be performed at the edge server. The compute-intensive parts/layers that the edge server cannot perform, can be offloaded to cloud server. The device only reports the sensing/perception data to the server and does not need to support AI/ML inference operations. The intermediate data are sent from the edge server to the cloud server. A proper split point needs to be selected for an efficient cooperation between edge server and cloud server.

•Mode c): Device-edge-cloud split inference

In this mode, an AI/ML inference operation or an AI/ML model is split over the mobile device, the edge server, and the cloud server. The compute-intensive parts/layers of an AI/ML operation/model can be distributed among the cloud and/or edge server. The latency-sensitive parts/layers can be performed on the device or the edge server.

Privacy-sensitive data can be left at the device. The device sends the intermediate data outcome from its computation to the edge server, and the edge server sends the intermediate data outcome from its computation to the cloud server. Two split points need to be selected for an efficient cooperation between the device, the edge server and the cloud server.

•Mode d): Device-device split inference

This mode provides a decentralized split inference. An AI/ML inference operation or model can be split over different mobile devices. A group of mobile devices can perform different parts of an AI/ML operation or different DNN layers for an inference task, and exchange intermediate data between each other. The computation load can be distributed over devices meanwhile each device preserves its private information locally.

•Mode e): Device-device-cloud/edge split inference

An AI/ML inference operation or model is first split into the device part and network part. Then the device part can be executed in a decentralized manner, i.e. further split over different mobile devices. The intermediate data can be sent from one device to the cloud/edge server, or multiple devices can send intermediate data to the cloud/edge server.

Although the above modes are described as in 5G system, the same principle can also be applied to the Next G networks. Distributed Learning and Federated Learning are considered as two basic methods.

Distributed Learning

In Distributed Learning, each computing node trains its own DNN model locally with local data, which preserves private information locally. To obtain the global DNN model by sharing local training improvement, nodes in the network will communicate with each other to exchange the local model updates.

Federated Learning

In Federated Learning, the cloud server trains a global model by aggregating local models partially trained by each end device. The most agreeable Federated Learning algorithm so far is based on the iterative model averaging. Within each training iteration, a UE performs the training based on the model downloaded from the AI server using the local training data. Then the UE reports the interim training results (e.g., gradients for the DNN) to the cloud server via uplink (UL) channels. The server aggregates the gradients from the UEs and updates the global model. Next, the updated global model is distributed to the UEs via downlink (DL) channels. Then the UEs can perform the training for the next iteration.

There are two ways for model collaboration in cross-domain AI: one is the collaboration between different models. Different domains process different sub-tasks with different models, while the input and output of the model are related, for example, the output of Model 1 serves as the input of Model 2. Another is the collaboration between the same model, where the model is divided into two parts from the middle layer based on data location or task type.

3.2.3 Computing resource

(1) Coordinated optimization of computing and communication

Designing cross-domain AI for computing and communication in wireless next generation networks requires a holistic approach that considers the unique characteristics of the network, the demands of different applications, and the needs of users.

State/data exchange architecture is an approach to optimize compute and communication across different domains in a network. This architecture involves exchanging data and state information between different domains, such as the edge and the cloud, to optimize resource allocation, reduce latency, and improve application performance. The proposed architecture is shown in Figure 3.4.

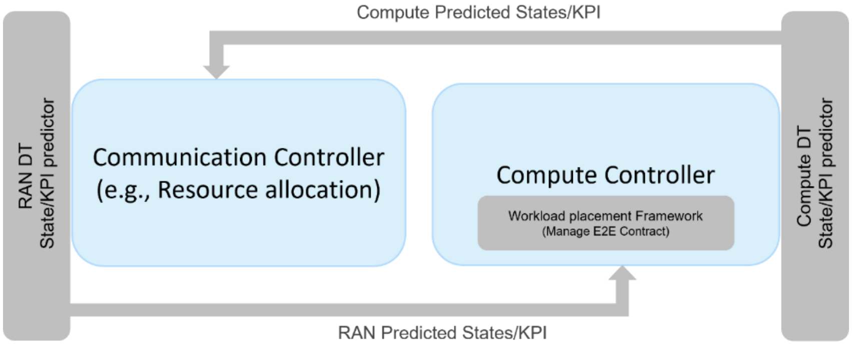
Compute Predicted States/KPI
RAN DT State/KPI predictor
Communication Controller (e.g., Resource allocation)
Compute Controller
Workload placement Framework (Manage E2E Contract)
Compute DT State/KPI predictor
RAN Predicted States/KPI

Figure 3.4 State assisted compute-communication co-design

As can be seen in this figure, Digital Twin RAN (DT-RAN) and DT-compute are used to predict domain specific KPIs and states to be shared with the other domain and be used by their own domain controller. DT-RAN and DT- compute can play a crucial role in predicting KPIs and states for joint compute and communication optimization in Next G networks. These digital twin technologies enable a virtual representation of the physical RAN and compute infrastructure, capturing real-time data and simulating its behavior.

By integrating the predictions from both DT- RAN and DT- compute, joint compute and communication optimization can be performed. The predicted KPIs and states guide resource allocation decisions, communication parameter adjustments, and computation offloading strategies.

Predicted KPIs from communication can be effectively utilized in the compute controller, while compute KPIs guide the decisions of the communication controller in a joint compute and communication optimization framework. In the compute controller, communication predicted KPIs provide valuable insights into the expected network conditions, such as latency, bandwidth availability, and congestion levels. This information can be utilized to make intelligent decisions regarding compute resource allocation and task scheduling. For example, if the communication predicted KPIs indicate high latency or limited bandwidth, the compute controller can prioritize offloading compute-intensive tasks to nearby edge servers or cloud resources to reduce the processing delay and improve the overall system performance. By considering the communication predicted KPIs, the compute controller can optimize the resource utilization, enhance application responsiveness, and ensure efficient task execution. Conversely, compute KPIs can play a crucial role in the decision-making process of the communication controller. These KPIs provide insights into the computational capabilities, workload demands, and resource utilization of the compute infrastructure. By integrating compute KPIs, such as CPU utilization, memory usage, and task execution time, the communication controller can dynamically adjust communication parameters to optimize the overall system performance. For instance, if the compute KPIs indicate high resource utilization or increased task execution time, the communication controller can allocate additional bandwidth or prioritize traffic to support the compute-intensive tasks, ensuring timely delivery of data and reducing processing delays. By considering compute KPIs, the communication controller can adapt the communication strategies, optimize resource allocation, and ensure efficient data transfer between compute nodes and end devices.

By leveraging the reciprocal relationship between communication predicted KPIs and compute KPIs, joint compute and communication optimization can be achieved. This integration enables a holistic view of the system, where communication decisions are informed by predicted compute states, and compute decisions are influenced by predicted communication conditions. This symbiotic relationship can facilitate efficient resource allocation, improved system performance, and enhanced user experience in Next G networks. By considering both communication predicted KPIs in the compute controller and compute KPIs in the communication controller, the joint optimization framework can dynamically adapt to changing network and compute conditions, leading to optimal allocation of resources, and improved overall performance.

The basic idea behind this architecture is to transfer data and state information between different domains in a timely and efficient manner. For example, in a wireless network, data can be transferred between the edge and the cloud to optimize resource allocation and reduce latency. The edge can process real-time data and perform local computations, while the cloud can perform more complex computations and store large amounts of data.

In this architecture, state information is used to keep track of the current state of the network, such as available resources, network traffic, and user preferences. This information is continuously updated and shared between different domains to optimize resource allocation, reduce latency, and improve application performance.

(2) Mapping AI tasks with computing resources

In the future, AI will be distributed across various locations in the network, including RAN, CN, and management systems, all of which can provide computing resources with different qualities. To better provide end-to-end intelligent services and improve the efficiency of computing resources, it is necessary to match the requirements of different AI tasks with the corresponding computing resources. Figure 3.5 shows the architecture performing this matching in the Cross-domain AI Controller.

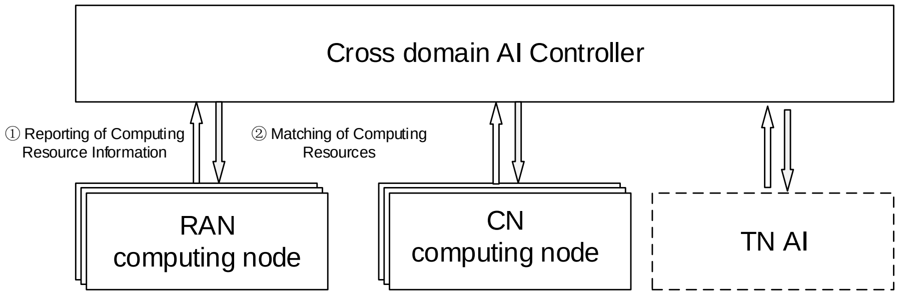
Cross domain AI Controller
① Reporting of Computing Resource Information
② Matching of Computing Resources
RAN computing node
CN computing node
TN AI

Figure 3.5 Matching of computing resources across different domains

Compared to the core network and management system, the computing resources on the RAN side are usually limited. However, its proximity advantage allows it to provide lower latency. Therefore, in resource matching for different AI tasks, it is necessary to consider factors such as latency requirements, task complexity, and available computing resources. For tasks that require real-time response, allocating some computing resources on the RAN side can provide lower latency. On the other hand, for computationally intensive tasks, it may be necessary to rely on the core network and management system to provide more powerful computing resources.

Therefore, in the matching of computing resources, it is necessary to consider the specific characteristics and requirements of the tasks, and comprehensively consider factors such as latency, computing power, and resource availability to achieve optimal resource allocation and performance optimization.

First, the Cross-domain AI controller in SMO needs to identify the requirements of AI tasks. For complex AI tasks, they can be decomposed into multiple parallel subtasks. Then, the cross-domain AI controller needs to obtain AI computing resource information from various nodes in the network, such as resource types, computing capacity, resource performance, load, etc. Finally, based on the task requirements and resource information, the AI tasks can be matched to different computing nodes for AI training and inference.

Besides, wireless network computing power is showing a trend of distributed deployment, and nodes such as base stations, network management, Multi-access Edge Computing (MEC), core networks, and data centers can deploy computing power, achieving a multi-level and three-dimensional distributed computing system at the cloud, edge, and device, to meet the compute requirements.

4 Conclusion

This research report presents a comprehensive overview of cross-domain AI. It first summarizes the current AI-related research and work of different standardization organizations, within the RAN, CN, management domain and digital twin domain. Then four application scenarios and technical requirements that need to be supported by intelligent collaboration in the future are identified. Finally, the report proposes a reference architecture for cross-domain AI, suggesting two potential technical solutions for end-to-end intelligent management and RAN-CN convergence. In the future, the 6G network architecture, cross-domain collaboration methodology, interface design, and management process to support AI collaboration can be studied in depth to enable a wide range of specific scenarios and use cases.

References

[1] O-RAN nGRG. RR-2023-03. Research Report on Native and Cross-domain AI: State of the art and future outlook.

[2] 3GPP TR 37.817 V17.0.0: "Study on enhancement for Data Collection for NR and EN-DC"

[3] 3GPP TR 38.843 V18.0.0: "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface"

[4] O-RAN.WG3. Near-Real-time RIC Architecture-v05.00.

[5] O-RAN.WG2. AI/ML workflow description and requirements-v01.03

[6] 3GPP TS 23.288 V18.4.0: "Architecture enhancements for 5G System (5GS) to support network data analytics services."

[7] 3GPP TR 28.809 V17.0.0: "Study on enhancement of Management Data Analytics (MDA)"

[8] 3GPP TS 28.104 V18.2.0: "Management Data Analytics (MDA)"

[9] 3GPP TS 28.105 V18.2.0: "Artificial Intelligence / Machine Learning (AI/ML) management"

[10] 3GPP TR 28.908 V18.0.0: "Study on Artificial Intelligence / Machine Learning (AI/ML)"

[11] ETSI, Zero Touch Network and Service Management (ZSM), Reference Architecture, ETSI GS 002 1.1.1, August 2019.

[12] 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.

[13] ONAP, "DCAE Architecture," [Online]. Available: https://docs.onap.org/projects/onap-dcaegen2/en/latest/sections/architecture.html

[14] D. Wang, R. Su and S. Zhang, "An Intent-based Smart Slicing Framework for Vertical Industry in B5G Networks," in Proc. 2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Xiamen, China, Jul. 2021, pp.389-394.

[15] ONAP, "R11 E2E Network Slicing use case," [Online]. Available: https://wiki.onap.org/display/DW/R11+E2E+Network+Slicing+use+case

[16] O-RAN. WG10. O-RAN Operations and Maintenance Architecture-v10.00.

[17] 3GPP R3-233756: "BLCR to 38423 - Support for AI-ML_CR0959r71"

[18] 3GPP TR 22.874 V18.2.0: "Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS"