GTI 5G Intelligent Network White Paper

GTI 5G Intelligent network

White Paper

Version:	V 1.0
Deliverable Type	□ **Procedural Document** √ **Working Document**
Confidential Level	**√Open to GTI Operator Members** **√Open to GTI Partners** □ **Open to Public**
Program	5G eMBB
Working Group	Network Working Group
Project	Project 6: Network Intelligence
Task	Intelligent Network Level (INL) Intelligent Network Architecture (INA) Intelligent Network Elements (INE) Intelligent Network Management (INM)
Source members	China Mobile, Huawei, CICT, Nokia, ZTE, Ericsson
Support members (in alphabetical order)	
Last Edit Date	
Approval Date	

Document History

Date	Meeting #	Version #	Revision Contents
Nov. 23,2020	29th GTI Workshop	V1.0	The first version of GTI 5G Intelligent Network Whitepaper. The standardization and industry status, the practical use cases and corresponding intelligent network level, architecture, function requirements on network elements and network management of 5G intelligent network are presented.

Table of Contents

1 Executive Summary

This is the first version of GTI white paper provides an overview of 5G Intelligent Network. It covers standardization status, practical use cases, intelligent network level, intelligent network architecture, intelligent network elements and intelligent network management.

After a brief review on intelligent network related activities in SDOs and industry parties, the whitepaper introduces a series of practical use cases of intelligent network for further study. The use cases can be categorized based on the dimension of full life cycle and main functional entities. Potential solutions and application with performance are introduced for each use case.

Intelligent network levels are beneficial for the industry to have a clear view on how to implement a fully intelligent network step by step. Here we summarize the framework approach for classification of intelligent network levels based on relevant standards. Evaluation on intelligent network levels of typical use cases are analyzed as well.

Intelligent network architecture is another important topic, in order to understand the impact of AI/ML on mobile network architecture, the whitepaper analyzes the framework architecture of intelligent network and implementation architecture from the practical uses case perspective.

Based on the analysis of use cases, levels and architecture of intelligent network, some general function requirements for intelligent network elements and intelligent network management are highlighted at last, which are derived from the implementation workflow and close-loop of intelligent network.

As it comes to 5G era, it is obvious that network has become an indispensable part of our lives. And network intelligence is widely considered as an important enabler for the network evolution with purpose of achieving promoted service performance and operational efficiency. This white paper is expected to provide helpful reference for any industry participates who are interested in and committed to promoting the development of network intelligence.

2 Reference

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

[1] GSMA, “AI in Network Use Case in China”, Oct 2019.
[2] ITU-T Y.3172: “Architectural framework for machine learning in future networks including IMT-2020”.
[3] ITU-T Y.3173: “Framework for evaluating intelligence levels of future networks including IMT-2020”.
[4] ITU-T Y.3174: “Framework for data handling to enable machine learning in future networks including IMT-2020”.
[5] ITU-T Y.ML-IMT2020-RAFR, Architecture framework for AI-based network automation of resource adaptation and failure recovery for future networks including IMT-2020.
[6] 3GPP TR 28.810: “Study on concept, requirements and solutions for levels of autonomous network”.
[7] 3GPP TS 28.100: “Management and orchestration; Levels of autonomous network”.
[8] 3GPP TR 28.812: “Telecommunication management; Study on scenarios for Intent driven management services for mobile networks”.
[9] 3GPP TS 28.312: “Intent driven management services for mobile networks”.
[10] 3GPP TS 28.535: “Management and orchestration; Management services for communication service assurance; Requirements”.
[11] 3GPP TS 28.536: “Management and orchestration; Management services for communication service assurance; Stage 2 and stage 3”.
[12] 3GPP TR 28.809: “Study on enhancement of Management Data Analytics (MDA)”.
[13] 3GPP TS 28.313: “Self-Organizing Networks (SON) for 5G networks”.
[14] 3GPP TR 28.861: “Study on the Self-Organizing Networks (SON) for 5G networks”.
[15] 3GPP TR 37.816: “Study on RAN-centric data collection and utilization for LTE and NR”.
[16] 3GPP TS 38.314: “NR; Layer 2 measurements”.
[17] 3GPP TS 38.300: “NR; Overall description; Stage-2”.
[18] 3GPP TS 37.320: “Minimization of Drive Tests (MDT); Overall description; Stage 2”.
[19] 3GPP TS 38.306: “NR; User Equipment (UE) radio access capabilities”.
[20] 3GPP TS 38.331: “NR; Radio Resource Control (RRC); Protocol specification”.
[21] 3GPP TR 23.791: “Study of enablers for Network Automation for 5G”.
[22] 3GPP TS 23.288: “Architecture enhancements for 5G System to support network data analytics services”.
[23] 3GPP TR 23.700-91: “Study on Enablers for Network Automation for 5G - phase 2”.
[24] ETSI GS ZSM 002: "Zero-touch Network and Service Management (ZSM); Reference Architecture”.
[25] ETSI GR ZSM 009-3: "Zero-touch Network and Service Management (ZSM); Closed-loop automation; Advanced topics”.
[26] ETSI GS ZSM 009-1: "Zero-touch Network and Service Management (ZSM); Closed-loop automation; Enabler”.
[27] ETSI GS ZSM 009-2: "Zero-touch Network and Service Management (ZSM); Closed-loop

automation; Solutions”.
[28] ETSI, ETSI GR ENI 007 V1.1.1 (2019-11): “ENI; ENI Definition of Categories for AI Application to Networks”.
[29] CCSA, “Technical report of telecommunication network planning application based on artificial intelligence”.
[30] CCSA, “Study on grading method for intelligent capability of mobile networks”.
[31] CCSA, “Technical specification for intelligent level of mobile network management and operation”.
[32] TMF: "A whitepaper of autonomous networks: empowering digital transformation for the telecoms industry", https://www.tmforum.org/wp-content/uploads/2019/05/22553-Autonomous-Networks-whitepaper.pdf.
[33] TMF: "A whitepaper of autonomous networks: empowering digital transformation for smart societies and industries", https://inform.tmforum.org/research-reports/autonomous-networks-empowering-digital-transformation-for-smart-societies-and-industries/
[34]Qing Zhang, “AI based intelligent recognition of 5g base station energy saving scenarios”, Aug 2019.
[35] Jian Feng Lei, “Research on network traffic prediction based on Neural Network“, May 2008.
[36] Zhi Rong Zhang, “Research on energy saving technology of 5G base station based on AI“, Oct 2019.
[37] China Mobile, “5G intelligent network white paper”, Dec 2018.

3 Abbreviations

Abbreviation	Explanation
3D	Three-Dimensional
3GPP	3rd Generation Partnership Project
AAU	Active Antenna Unit
AGV	Automated Guided Vehicle
AI	Artificial intelligence
API	Application Programming Interface
BTS	Base Transceiver Station
CA	Carrier Aggregation
CE	Cell Edge
CM	Configuration Management
CQI	Channel Quality Indication
DL	DownLink
DT	Drive Test
E2E	End to End
eMBB	enhanced Mobile Broadband
EMOS	Electric Operation Maintenance System
EMS	Network Element Management System
eNB	E-UTRAN NodeB
EOMS	Electric Operation Maintenance System
E-RAB	Evolved Radio Access Bearer
FP	Frequent Pattern
GANA	Generic Autonomic Networking Architecture
GTI	Global TD-LTE Initiative
HO	HandOver
INE	Intelligent Network Elements
INL	Intelligent Network Level
INM	Intelligent Network Management
KNN	K-Nearest Neighbor
KPI	Key Performance Indicator
LB	Load Balance
LTE	Long Term Evolution
MDT	Minimization of Drive Tests
MEC	Multi-Access Edge Computing
MIMO	Multiple-input multiple-output
ML	Machine Learning
mMTC	massive Machine-Type Communications
MR	Measurement Report
MU-MIMO	Multi-User Multiple-Input -Multiple-Output
NE	Network Element

NMS	Network Management System
NR	New Radio
NWDAF	Network Data Analytics Function
O&M	Operation and Maintenance
OPEX	Operating Expense
PDU	Protocol Data Unit
PM	Performance Management
PRB	Project Review Board
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RCA	Root Cause Analysis
RF	Radio Frequency
ROI	Return On Investment
RRC	Radio Resource Control
RRM	Radio Resource Manage
RRU	Radio Remote Unit
RSRP	Reference Signal Receiving Power
SCTP	Stream Control Transmission Protocol
SDO	Standards Development Organizations
SON	Self-Organizing Network
TTI	Transmission Time Interval
UE	User Equipment
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communication
VNF	Virtual Network Function
VR	Virtual Reality

4 Introduction

This whitepaper mainly focuses on the 5G Intelligent Network. Combined with standardization status, the whitepaper analyses a serious of practical use cases, intelligent network levels, intelligent network architectures of the representative use cases, and general function requirements on intelligent network elements and intelligent network management are presented as well. This Whitepaper is expected to help people make a comprehensive understanding of intelligent network and its industry. It is also expected to be helpful on promoting the development of network intelligence.

Sincere thanks to all the contributors and the supporters for their hard work in writing this whitepaper. Respectfully, the task leaders and contributors of each chapter are listed as following.

- **Chapter 1 Executive Summary**
  China Mobile
- **Chapter 2 Reference**
  China Mobile, CICT, Ericsson, Huawei, Nokia, ZTE
- **Chapter 3 Abbreviations**
  China Mobile, CICT, Ericsson, Huawei, Nokia, ZTE
- **Chapter 4 Introduction**
  China Mobile, CICT, Ericsson, Huawei, Nokia, ZTE
- **Chapter 5 Standardization and Industry Status**
  China Mobile, Huawei
- **Chapter 6 Use cases**
  CICT, Huawei, Nokia, ZTE, Ericsson, China Mobile
- **Chapter 7 Intelligent Network Level**
  Huawei, CICT, Ericsson, China Mobile
- **Chapter 8 Intelligent Network Architecture**
  Nokia, Ericsson, Huawei, CICT, ZTE, China Mobile
- **Chapter 9 Intelligent Network Elements**
  ZTE, Nokia, Ericsson, CICT, China Mobile
- **Chapter 10 Intelligent Network Management**
  Huawei, CICT, Ericsson, China Mobile

Special thanks to the following contributors for writing the whitepaper.

- **China Mobile**
  Dr. Xi Cao, Ms. Xiaojia Song, Ms. Yanping Liang, Dr. Lin Zhu, Ms Ruoyi Liu, Mr. Peng Zhao, Mr. Shaofan Chen, Ms. Xing Wang, Mr. Li Yu, Mr. Xiangyang Yuan, Dr. Junlan Feng
- **CICT**
  Ms. Xiufeng Wang, Ms. Xuan Zhang
- **Ericsson**
  Ms. Xiaomei Liu, Mr. Bo Guan, Mr. Liyu Liu
- **Huawei**

Ms. Danni Chen, Mr. Ruiyue Xu, Mr. Junjie Zhang, Mr. Yuanxian Tang, Mr. Chunhong Yuan, Ms. Lan Zou, Mr. Li Lin

- **Nokia**
  Mr. Zhuyan Zhao, Ms. Fangfang Gu
- **ZTE**
  Mr. Xinjian Jiang, Mr. Pan Li, Mr. Wanteng Zhai, Ms. Hong Feng, Mr. Fei Ye, Mr. Honghui Kang, Mr. Xuefeng Lu

This is the first version of the whitepaper, it will be continuously updated according to the research and development progress.

5 Standardization and Industry Status

5.1 Motivation and Overview

As mobile communication network evolves to 5G, it is obvious that the network has become an indispensable part of our lives. While being the enabler, mobile network itself is evolving into the intelligence era with multiple application scenarios, features, services and operation requirements for the intelligent network. Technologies including artificial intelligence (AI) are expected to be introduced to enable autonomous networks in the areas of network planning, deployment, operation, optimization, service deployment, assurance, etc.

Most of the standards development organizations (SDO), e.g. ITU-T, 3GPP, ETSI, CCSA, are taking actions to study and develop standards for autonomous networks, topics related to autonomous networks are discussed in different working groups.

Industry bodies such as GSMA, TM Forum, Global TD-LTE Initiative (GTI) etc. are working to promote autonomous networks. GSMA proposed that the automatic network operation capability will become the indispensable 4th dimension of the 5G era together with eMBB, mMTC, and URLLC, and become one of the most important driving forces for 5G service innovation and development. [1]

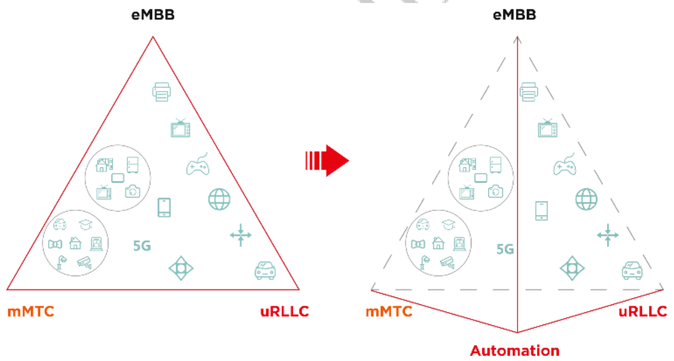

Figure 5-1 Network automation is the 4th dimension of 5G networks [1]

Main activities about autonomous networks in the SDOs and industry parties are introduced briefly in the following.

5.2 Activities in ITU

In ITU-T, Study Group 13 (SG13) has led the ITU standardization work on next generation networks and now caters to the evolution of NGNs, while focusing on future networks and network aspects of mobile telecommunications. A focus group on Machine Learning for Future Networks including 5G (FG-ML5G) had been set up for machine learning for future networks, which includes interfaces, network architectures, protocols, algorithms and data formats. The topic related to autonomous networks came into study since 2017 and recommendations ITU-T Y.3172 (Architectural framework for machine learning in future networks including IMT-2020) [2], ITU-T Y.3173 (Framework for evaluating intelligence levels of future networks including IMT-2020) [3], ITU-T Y.3174 (Framework for data handling to enable machine learning in future networks including IMT-2020) [4], etc. have already published. Recommendations about AI-based network autonomous, e.g. ITU-T Y.ML-IMT2020-RAFR [5] are in draft stage.

5.3 Activities in 3GPP

Autonomous network comes into the sight of 3GPP since 4G era, the topics mainly focused on Self-Organizing Network (SON), Minimization of Drive Tests (MDT), etc. In 5G era, 3GPP makes more efforts on the standardization for autonomous networks.

- 3GPP SA WG5 started the study item "Study on autonomous network levels" [6] from August 2019 and published in September 2020, which output the concept, dimension, framework and typical use cases for classification of autonomous network level. In June 2020, a new work item "Autonomous network levels" [7] was approved and started. The main objective of this work item is to deliver standard specifications for defining concept and framework for autonomous network level, and providing requirements for corresponding autonomous network enabler related standard features for different autonomous network levels.

- Since Release 16, 3GPP SA WG5 has started a series of standard projects to continuously promote the autonomous network, which cover the entire mobile network lifecycle (including planning, deployment, maintenance, and optimization phases). Examples of these projects are "Intent driven management service for mobile network" [8] [9], "Closed loop SLS Assurance" [10] [11], "Study on enhancement of Management Data Analytics Service" [12], and "Self-Organizing Networks (SON) for 5G networks" [13] [14].

- 3GPP SA WG2 and 3GPP RAN WG3 have also started some works on topics related to network automation and intelligence. Since Release 16, 3GPP RAN WG3 started the "RAN-centric Data Collection and Utilization" research topic [15] and "SON/MDT support for NR" standard project [16-20], and researched and defined wireless data collection and application oriented to network automation and intelligence. Since Release 16, 3GPP SA WG2 has started the "Enablers for Network Automation for 5G" standard project [21-23]. The objective is to define the introduction of Network Data Analytics Function (NWDAF) on the 5GC network layer to implement control-plane data analysis for 5GC.

5.4 Activities in ETSI

The study on autonomous network is active in ETSI, there are several groups working on relevant topics of autonomous networks: ENI, NFV, OSM, MEC, F5G, TC INT AFI and ZSM.

- Generic Autonomic Networking Architecture (GANA) is studying TC INT AFI.

- ETSI ZSM targeted to deliver E2E solution for Zero-touch network and service management and orchestration, and ZSM framework is designed for closed-loop automation and optimized for data-driven machine learning and AI algorithms. ZSM published Zero-touch network and service management reference architecture specification [24] and started discussions on three closed loop topics related projects: "closed-loop automation: Advanced topics" [25], "closed-loop automation: Enablers" [26], and "closed-loop automation: Solutions for automation of E2E service and network management use cases" [27] in June 2019. The goal is to deliver the use case, requirements and solutions for the E2E closed loop automation including automation-related policies and intent interfaces to implement interaction between closed loops in E2E management domains and management domains.

- In November 2019, ETSI published the report ETSI GR ENI 007[28]: ENI definition of categories for AI application to networks which defines various categories for the level of application of AI techniques to the management of the network, going from basic limited aspects, to the full use of AI techniques for performing network management.

5.5 Activities in CCSA

As one of the most influential SDOs in the field of communication in China, CCSA began the standard works on autonomous networks from 2010s, and the items are mainly set up in TC1, TC5 and TC7, including use cases, architecture, data handling, levels of autonomous network, management requirements etc.

- CCSA TC1 started the research project [29] on telecommunication network planning application based on artificial intelligence from July 2018. The research is focusing on network planning based on AI. The output report will mainly cover network planning needs, network architecture, data training, network evolution.
- CCSA TC5 started the research project [30] on intelligence levels of mobile networks in November 2018 and finished in August 2019. The output of the research covers the methods for evaluating intelligence levels for mobile networks, and typical use cases for classification of intelligence level and potential relationships between the mobile network architecture and intelligence levels.
- CCSA TC7 started the standard specification "Technical requirements for intelligence levels of mobile network management and operation"[31] in January 2020. The work target to define the concept, measurable evaluation method, typical use case and requirements for intelligence levels of mobile network management and operation.

5.6 Activities in Industry Parties

Industry bodies such as GSMA, TM Forum are actively taking actions to explore and promote the collaboration of autonomous network topics among the SDOs, operators, vendors and any other industry participants.

In GSMA, AI & Autonomous is one the topic of the “Future Network”. In June 2019, the first GSMA Global AI Challenge was held and the challenge investigated three specific areas: connectivity in rural areas, mobile energy efficiency and enhanced services in urban areas. In June, at the AI in Network Seminar for Mobile World Congress Shanghai 2019, it is called on the entire industry to focus on and contribute to the key applications of AI in the mobile network, and jointly build the 5G era for the intelligent autonomous network in the workshop. In October 2019, GSMA has published “AI in network use cases in China” [1].

In TM Forum, several workshops of autonomous have been held since 2019 and Autonomous Networks Project (ANP) was established in August 2019. Since 2019, there are three whitepapers are published officially: AN Whitepaper 1.0, IG1193 Vision & Roadmap v1.0, IG1218 Business requirement & architecture v1.0. [32] This year, AN Whitepaper 2.0 [33], business requirements & architecture v1.1, technical architecture, demo of Catalyst projects, user stories/use cased, etc. are now on going.

5.7 Summary

As the development of technologies and the evolution of networks, intelligent network will be an important enabler for the future networks. In order to promote the industry, SDOs and industry parties are taking activities to form a unified understanding and continuously clarify the concept.

6 Use Cases

6.1 Introduction

Operators, vendors and third-parties have already begun to explore intelligent networks, and a number of good practices and use cases have emerged.

This whitepaper categorizes the intelligent network use cases based on the dimension of full life cycle i.e. network and service planning, network and service deployment, network and service maintenance, network and service optimization, and the dimension of main functional entities i.e. intelligent network element and intelligent network management. And for each use case, background, solution overview, application and performance are presented.

6.2 Classification for Use Cases

6.2.1 Full Life Cycle Dimension

- **Network and Service Planning:** processes of designing and delivering new or enhanced network or service based on the business, market, product and customer service requirements.
- **Network and Service Deployment:** processes of allocation, installation, configuration, activation and verification of specific network and service.
- **Network and Service Maintenance:** processes of monitoring, analyzing and healing of the network and service issue.
- **Network and Service Optimization:** processes of monitoring, analyzing and optimization/assurance of the network and service performance.

6.2.2 Functional Entity Dimension

- **Intelligent Network Element:** the close loop of intelligent function is mainly deployed within and completed by the entity of network element (with network management system's control).
- **Intelligent Network Management:** the close loop of intelligent function is mainly deployed within and completed by the entity of network management system (with network elements' assistance).

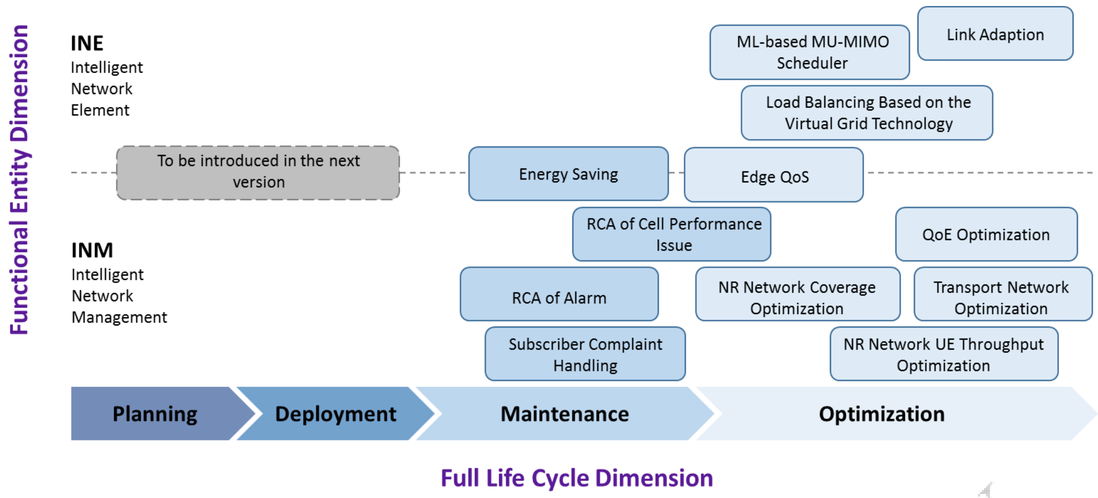

Figure 6-1 Classification of use cases

Considering the current development status of the industry, this version of white paper, mainly focuses on the use cases of network and service optimization and maintenance. Use cases of network service planning and deployment would be introduced in the next version.

6.3 Use Cases of Network and Service Maintenance

6.3.1 Energy Saving

6.3.1.1 Background

As operators' network energy consumption keeps increasing, reducing the energy consumption of main equipment is key to energy saving. Reducing the power consumption of main equipment of wireless sites has become the top priority for all. For a typical carrier, the power consumption of wireless sites accounts for about 45%, and the power consumption of wireless base stations as main equipment accounts for 50%. In the power consumption of a wireless base station, the power consumption of radio remote units (RRUs) accounts for a large proportion, and that of the power amplifiers in the RRUs also accounts for a large proportion. In actual networks, traffic has obvious tidal effect in most cases. When the traffic is light, the base station is still running, which causes a great waste of energy.

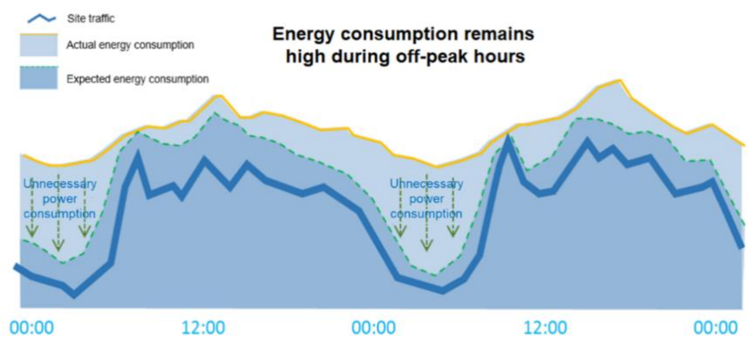

Figure 6-2 Challenges facing traditional energy saving

Reducing unnecessary power consumption is a key measure of energy saving but is faced with many challenges. The network traffic volume varies greatly during peak and off-peak hours. The equipment keeps running, and the power consumption is not dynamically adjusted based on the traffic volume. As a result, a waste of resources is caused. The capability of "zero bits, zero watts" needs to be constructed. However, in a typical network, the features of different scenarios vary greatly. How to automatically identify different scenarios and formulate appropriate energy saving policies becomes the key to energy saving.

- Business district: high requirements on user experience, obvious tidal effect, and light traffic at night.
- Residential area: high requirements on capacity, heavy traffic in a whole day, and no obvious traffic fluctuation.
- Suburban area: low requirements on capacity, light traffic, sparse sites, and long site coverage distance.

To meet the need of environmental-friendly development featured in low-carbon, energy saving and emission reduction and the requirement of cost reduction from telecom operators, the contradiction between the increasing communication data service volume and high energy consumption needs to be resolved while ensuring the development of 5G services. Therefore, energy saving technologies have always been a hot topic in the industry, and equipment energy saving has always been an important direction of research. As 5G technologies become mature, it is urgent to accelerate the commercial application of energy-saving solutions for 5G equipment.

6.3.1.2 Solution Overview

Based on network-level AI-based intelligent energy saving policy management and site energy saving scheduling control, the mobile network energy saving solution implements network scene adaption, one site one policy, and multi-network collaboration for intelligent base station energy saving management. This maximizes network energy saving benefits while ensuring stable network performance, and achieves the optimal balance between energy consumption and KPIs. The overall solution is as follows:

- The system obtains data on the live network, including engineering parameters, MRs, and weather data.

- Based on big data analysis, the system uses AI technologies to automatically identify network energy saving scenarios, predict network traffic trends, such as traffic busy/idle hours and areas and traffic/energy consumption trends, and identify multi-cell co-coverage, and automatically generates energy saving policies.

- The system automatically delivers energy saving policies and implements network-level AI-based intelligent energy saving policy management and coordinated management and control of site energy saving scheduling.

  - The network-level AI-based energy saving algorithm is used to implement automatic precise energy saving feature enabling and energy saving parameter optimization with lossless performance based on different network scenarios/models, base station configurations, and networking modes (multi-frequency networking and 2G/3G/4G/5G multi-RAT networking). In addition, the solution implements one site one policy and multi-site collaboration to quickly and efficiently start network-wide energy saving.

- Precise energy saving scheduling control (such as carrier shutdown and power adjustment) is implemented for sites under the control of network AI.

• Real-time monitoring of impact on network KPIs and energy saving benefits is implemented to achieve manual visualization and management of energy saving benefits on mobile networks.

6.3.1.2.1 Energy Saving Scenario Identification

For traditional energy saving, due to the diversity of scenes across the network and the large differences in the characteristics of the scenarios, manual operation cannot effectively identify the different scenarios and can only use the same set of energy saving policies for different scenarios.

AI-based scenario identification can intelligently identify and label the scenario of each cell based on network coverage, users, resources, and other characteristic data.

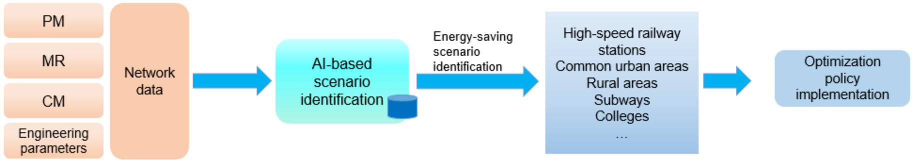

Figure 6-3 Scenario identification for energy saving

Scenario identification includes the following two levels:

The first level can identify the coverage scenario of a cell, for example, whether it is a high-speed railway, common urban area, rural area, subway, large stadiums, colleges, shopping malls, or office buildings. After scenario identification, for different scenarios can be combined with the characteristics of various energy-saving technologies to preset the recommended energy-saving solutions by scenario, as well as adapting to find the best recommended energy-saving solutions. The specific energy saving technologies include carrier shutdown, channel shutdown, symbol shutdown, cell shutdown and so on.

The second level is based on network topology data (for example, engineering parameters and configuration data of the cell), measurement reports, and handover indicators, the system can identify whether a cell is the coverage/capacity-layer cell, overlapping coverage degree of a cell pair, and whether a cell has the same-coverage cell. The result can be used as the input of intra-RAT/ inter-RAT collaborative energy saving.

The first step to implement power saving policy for a specific cell is to intelligently identify the scenario of the cell. During scenario identification, each base station can identify the feature data such as topology information, uplink/downlink measurement report, service characteristics, user level information and resource occupation distribution of each cell. The AI algorithm can use such classical machine learning algorithms as K-means clustering algorithm, KNN algorithm, decision tree and logic regression for scenario prediction and classification.

Scenario identification is supported on both the EMS and the NE. When it is implemented on the NE side, more user-level and service-level data is provided to improve the accuracy of identification, but the NE side has the AI data storage and computing capabilities.

6.3.1.2.2 Traffic Prediction

Based on historical network data, for example, time, cell traffic statistics, neighbor cell relations, handover data, holidays, and major events, AI modeling is performed at the cell, cell cluster or region level to predict the load flow direction and load level of a cell or cell cluster in the next few hours. Based on the load prediction result, it is used to accurately understand the occurrence time and duration of low load. In this way, intelligent energy saving can be performed on cell. In addition, based on historical handover and load migration information, the prediction model can predict the neighbor cells to which the load of a cell is transferred when the cell enters energy saving state, and then make adaptive settings for the energy saving start and end time of the neighbor cells and the load threshold. [34]

Load prediction can be supported on both the EMS and the NE sides. However, if it is implemented on the NE side, load prediction can be implemented based on whether each cell is in energy-saving status. In this way, the accuracy of neighbor cell load prediction can be improved.

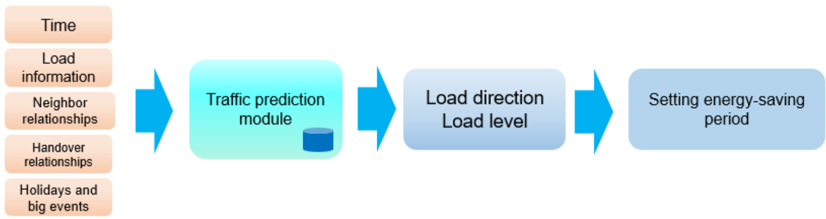

Figure 6-4 Traffic prediction for energy saving

The AI approach can accurately predict the effective time period for energy saving applications, thereby reducing the impact on performance KPIs caused by irrational energy saving time period configurations in manual configurations.

6.3.1.2.3 Intra-RAT/ Inter-RAT Collaborative Energy Saving

The objective of collaborative energy saving is to select compensable cells (or groups of cells) simultaneously when selecting energy-saving cells so as to form a group of collaborative cells. Specifically, the gNodeB selects a coordination cell group in accordance with the same-coverage cell, overlapped coverage degree between cells and neighbor cells, coverage scenario, and the real-time load prediction result of each cell. After appropriate energy saving methods are enabled, the handover parameters of the energy-saving cell and neighbor cells can be intelligently adjusted to predict the load guarantee capability of the neighbor cells when a cell is in energy saving status. [35]

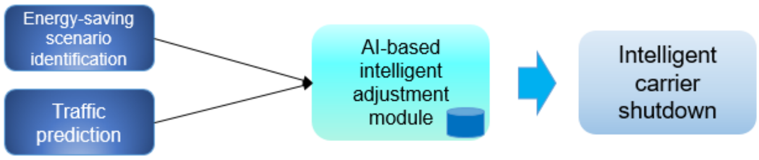

Figure 6-5 Intelligent carrier shutdown for energy saving

For intra-RAT coordination, the load prediction results of overlapping coverage and candidate supplementary cells are used to determine whether the target cell can be used as a compensation cell. For inter-RAT coordination, the support of candidate cells for the QoS level of the energy-saving cell is added. The following describes how to select a compensation cell.

Channel shutdown means that when the cell load is low, the shutdown of some transmission channels of the local cell which may affect edge coverage and the download rate. For a cell with channel shutdown is enabled, it is necessary to intelligently evaluate whether the coverage of edge users and downlink services of the cell can be serviced by adjacent cells within the radio mode when the cell is in channel shutdown status.

Carrier shutdown is applicable to the scenarios where multiple layers of networks are covered, and the traffic tidal effect is obvious, such as high-speed railways, universities, subways, and large stadiums. Basic coverage can be guaranteed by the coverage layer cells during idle times, and the capacity layer cells' carriers are turned off for energy saving. Therefore, carrier shutdown must ensure that there are two or more carriers covered in the same sector. This action cannot be performed if there is only one carrier covered in a sector, because once it is performed, the entire sector will be without signal. Therefore, before carrier shutdown, it is necessary to intelligently analyze the same- coverage situation.

Note: To determine energy saving, you need to select energy saving objects and load migration objects on the NE side in real time. Therefore, it is recommended that you select energy saving objects on the NE side.

6.3.1.2.4 Optimization of Energy Saving Policy Parameters

At present, the trigger thresholds for various types of energy saving in the energy saving policy (downlink PRB utilization, number of RRC users, etc.) are mainly set based on manual experience, and the thresholds are not optimized according to the actual energy saving effect.. For energy saving, if the shutdown policy is loose (for example, if the downlink PRB utilization threshold, which triggers entry into the energy-saving state, is set high, it is easier for the cell to enter the energy-saving state, , which means that the policy is loose), the energy saving effect is better. However, if the policy is too loose, the cell can easily enter energy saving state, which may reduce the service quality and traffic requirements of users. Otherwise, it is difficult for the cell to enter the energy-saving status, and the energy-saving effect is poor.

Energy saving policy parameter optimization aims at the threshold for triggering energy savings (for example, DL PRB usage). The purpose is to iteratively evaluate the energy saving effect and KPI performance of a cell/cell group under various triggering thresholds, and obtain the energy saving triggering threshold with the best KPI and energy saving effect. The recommended energy saving policy threshold for a cell/cell group is obtained. That is, the inflection point between the energy saving policy/load threshold and performance/energy saving effect is obtained to maximize the energy saving effect. The comprehensive score of KPI and energy saving effect in the iteration process can be defined in the form of target function, for example, α * (1- call drop rate) + β * access success rate + γ * normalized average throughput + μ * normalized energy consumption, where α, β, γ, μ are the weights of each item. Specific definitions are recommended based on the operator's attention to the performance index and the impact of the energy saving action on the network.

The system can automatically optimize energy-saving policies and load thresholds based on traffic prediction and reinforcement learning, and implement online iterative optimization to optimize the shutdown duration without affecting KPIs.

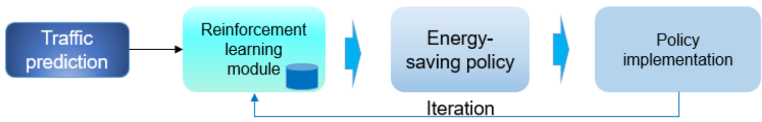

Figure 6-6 Energy-saving policy optimization

During prediction modeling, the KPIs of key network indicators need to be monitored, and the current prediction models need to be fed back in accordance with the changes of KPIs to achieve the most advantages of energy saving and system performance.

Note: The parameters of an energy saving policy can be optimized through rough adjustment of outer-loop parameters on the EMS side, and fine adjustment of inner-loop parameters and individual parameters can be performed on the NE side to assist in parameter optimization.

6.3.1.2.5 Enhanced Symbol Shutdown Based on Intelligent Scheduling of Wireless Resources

By intelligently and dynamically adjusting cell-level uplink and downlink resources, the system optimizes the service symbol resources allocated to users under the condition that service delay and service level are met. With light network load, the system maximizes the ratio of symbols in energy saving status and improves energy saving efficiency.

The AI algorithm predicts service distribution and load in future based on historical user load/service analysis, current user service type analysis, and service arrival analysis. As the input of symbol resource scheduling, the AI algorithm optimizes symbol resource scheduling without affecting user experience and enables idle symbols to enter energy-saving status in a timely manner.

Note: This module can only be supported on the NE side.

6.3.1.3 Application and Performance

In typical network configurations, the power consumption of base stations can be reduced by 10%–15%, and the emission of about 2 million kg carbon dioxide can be avoided for every 1000 base stations in one year.

Operator in China applies AI technologies and automation capabilities to base station energy saving. The RAN element management system (EMS) can automatically identify different scenarios and optimize energy saving policies for different networking modes and loads, maximizing network energy saving benefits while ensuring KPIs. Energy saving solution is applied more than 11,000 cells in the entire province. The overall energy consumption is reduced by 13.59%. The average shutdown duration is 9.88 hours, which increases by 57% compared with that when the feature is manually enabled. The tidal effect is obvious in office buildings, business centers, large stadiums, suburban areas, and county-level areas. The average energy consumption is reduced by 16.88%.

On the basis of deploying energy saving functions at multiple layers such as base station software and hardware, and terminals, AI technologies are used to intelligently deploy scenario-based and cell-level refined energy saving policies, minimizing network energy consumption while ensuring stable KPIs.

The AI-based intelligent energy-saving technology collects historical and spatial feature data of each cell on the network to analyze the change rule of radio resource utilization, automatically identifies the coverage characteristics of cells and fully considers the network coverage, UE distribution, and scenario characteristics based on the prediction and evaluation results of coverage scenarios and traffic variables. In this way, the energy-saving policy can be self-adaptive or selected based on the operator's policy.

The following figure shows the online energy saving solution for wireless networks based on the layer- and domain-based principle.

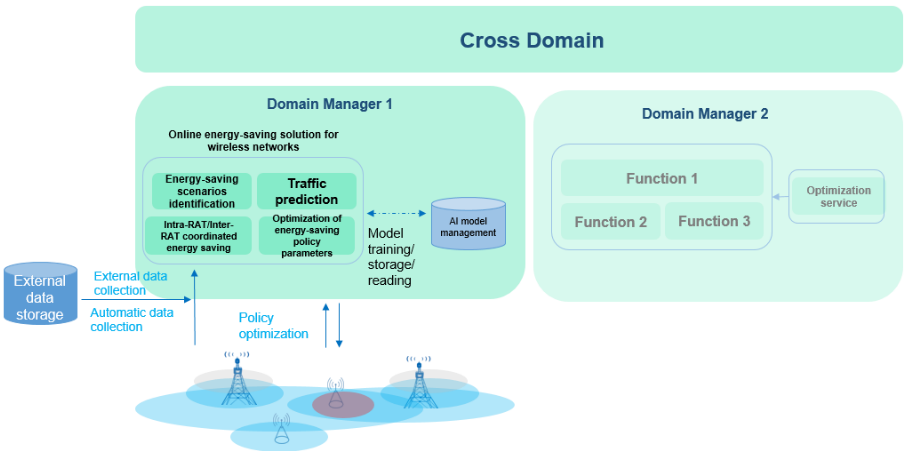

Figure 6-7 Architecture of the online energy-saving solution for wireless networks

The online energy-saving solution consists of four sub-functions:

- Energy-Saving scenario identification

- Traffic prediction
- Intra-RAT/Inter-RAT coordinated energy saving
- Optimization of energy-saving policy parameters

The AI-based wireless network solution can implement energy-saving policies such as cell sleeping and carrier shutdown through inter-RAT and intra-RAT coordinated management based on cell scenarios and energy-saving time, cutting energy consumption by 15%.

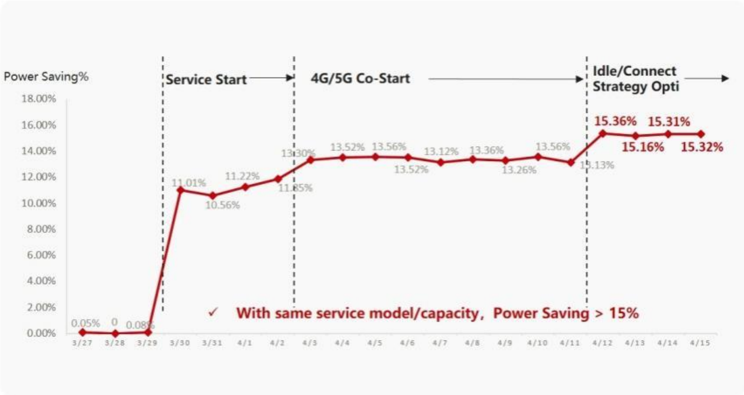

Figure 6-8 Simulation result of online energy-saving solution

6.3.2 Root Cause Analysis of Alarm

6.3.2.1 Background

With the rapid development of communication network in recent years, its scale has been quite large. In the network, there will be alarm information every day, and the amount of these information data is huge, and there are many sudden failures. When the network equipment fails and causes alarm, the equipment associated with it will also cause corresponding faults, and generate a large number of alarm information in a short time. As a fault often causes multiple alarm events, the equipment and business process related to the fault will send out relevant alarm information. At the same time, the alarm information caused by multiple faults will be superimposed together, which will submerge the real alarm information, which makes fault identification very difficult.

The rapid recovery of network fault is the basis to ensure the stable operation of the network. The traditional fault handling method is completed by manual analysis through a combination of network alarm, operation status and log data for manual analysis, and rely on reliable expert experience to achieve fault analysis and recovery. Analysis efficiency and screening effect were low in time dimension and regional dimension.

In 5G the network, combined with big data analysis and machine learning algorithm, the fault analysis experience can be informationized and modeled. Through multi-dimensional analysis of alarm information, network performance, operation log, etc., the association model that is difficult to be found manually can be mined out to form a precise root cause analysis system, which helps to improve the efficiency and success rate of fault analysis and recovery in the whole network. At the same time, by accumulating and sharing a large number of case data in the system, the fault prediction based on network operation and maintenance can be realized, and timely treatment and prevention can be obtained before the fault occurs, so as to improve the stability of network operation.

6.3.2.2 Solution Overview

Alarm root cause analysis is to use machine learning algorithm to train alarm association rules from massive alarm information, and combine with expert rule bases to form alarm diagnosis model bases. The existing network alarm information is diagnosed by matching rules, and the root cause alarm and derivative alarm are analyzed. The root cause alarm is accurately dispatched to improve the efficiency and success rate of analysis and recovery. [36]

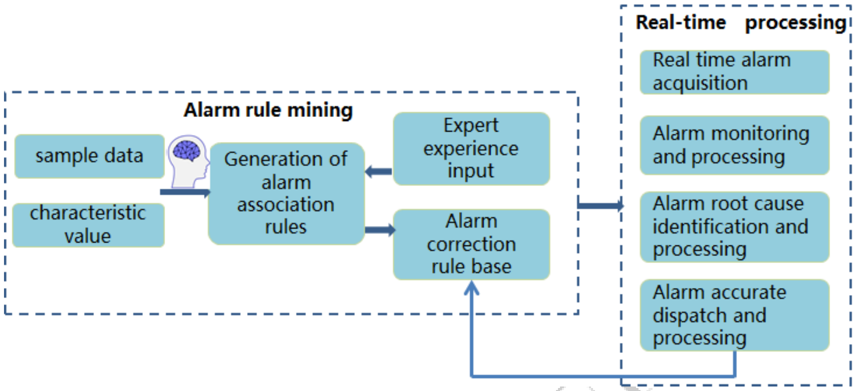

Figure 6-9 Solution chart

Alarm root cause analysis scheme is divided into two stages: alarm rule mining stage, real-time alarm analysis and processing stage.

The purpose of alarm rule mining stage is to analyze the big data based on historical alarm data, and obtain the relationship between alarms (for example, it can be based on Apriori and FP [37] In this stage, offline processing can be used to analyze and mine historical data, and it is not required to be real-time.

In the alarm analysis and processing stage, the purpose is to analyze and process the real-time alarms in the network based on the association rules in the rule database, and identify the source alarms and derived alarms. In this stage, online processing is used to process real-time alarm, which requires real-time performance.

6.3.2.3 Application and Performance

The current alarm data is monitored in real time in the existing network. When a new alarm is received, it is matched with the alarm association rule base to analyze the alarm root cause and derived alarm. Then, according to the root cause and the derived alarm relationship, the high-efficiency alarm management, such as "alarm elimination", "alarm merging" and "associated alarm dispatch", are implemented. It is helpful to improve the efficiency and success rate of fault analysis and recovery in the whole network. It is expected that the efficiency of alarm troubleshooting can be improved 80%.

6.3.3 Root Cause Analysis of Cell Performance Issue

6.3.3.1 Background

For the increasingly large and complex multi-layer and multi-standard wireless communication network, whether it is network planning, maintenance, or network optimization, operators and equipment vendors are facing new opportunities and challenges. To deal with these new challenges, Artificial intelligence is one of our most important solutions for managing modern networks.

For increasingly dense multi-layer networks, the traditional way is mainly for network optimization background engineer to check a single KPI through a variety of network optimization tools, and then combine one or more KPIs with the experience of senior network optimization engineers to analyze them, and pre-defined check rules or threshold values, first find out the problem cell, and then analyze the problem cell with more relevant KPIs and MR information to find out the cause of the problem, and then give the corresponding solution optimization plan for optimized implementation and verification . This traditional way of dealing with the problem of dense and massive communities in large modern cities requires more and more network optimization engineers to deal with it.

The method of using machine learning algorithms to autonomously discover problems in massive data and quickly identify and classify problems can significantly improve the efficiency of network optimization work. The transformation and skill upgrade of network optimization personnel and the application of AI modules make fewer network optimizations. The network engineers can handle more and more complex network problems.

6.3.3.2 Solution Overview

Combined with the actual network cells' performance data (KPI), and make full use of the experience and skills of AI experts to complete the exploration and selection algorithms by machine learning on big data platform, take full advantage of the distributed computing power of AI big data platform, complete Iterative development and model training and verification of application core modules. And quickly integrate and docker with the input data of the existing network, and through the deployment of the containerized platform, it can be directly applied to the rapid classification and root cause analysis of the daily network optimization problem cells.

Through the selection of about 100 KPIs from a large number of KPIs, and through multiple iterations of rigorously trained machine learning algorithms, the automatic aggregation and classification of 12 major types of network problems is quickly realized, and the intelligent root cause analysis module is further used to provide various problems Network root cause. Significantly reduce the workload of on-site network optimization engineers, improve the efficiency of on-site optimization, and quickly maintain the optimal state of the entire network cell performance.

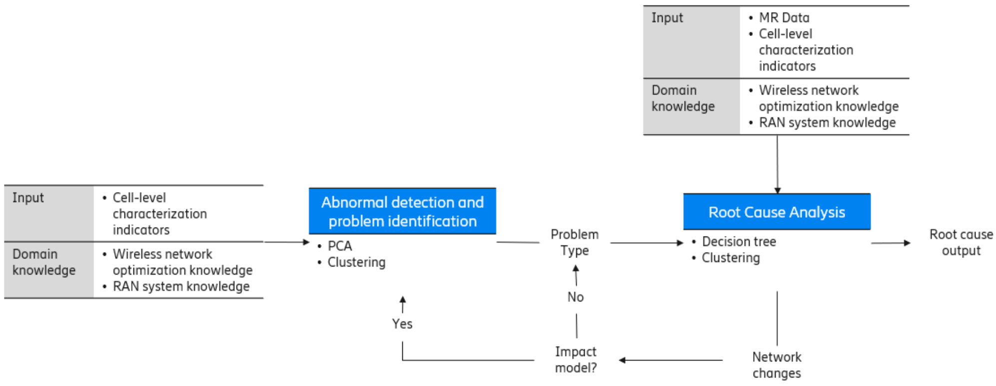

Figure 6-10 Root cause analysis system flow chart

The intelligent root cause analysis (RCA) solution for automatic network problem cells, on the one hand, realizes the identification of intelligent problem cells in multi-mode networks in one or more cities and gives the root causes of the main problems, and realizes automation through on-site deployment through containerized solutions; On the other hand, it also significantly simplifies the investigation and analysis of daily network optimization engineers' problem communities.

6.3.3.3 Application and Performance

Through the application and on-site verification in typical cities (10000+ cells), the AI module can quickly, accurately and intelligently identify problem cells and give the root causes of problems in the cell.

- 3 minutes to process more than 10k+ cells
- For the same workload, it takes two weeks of labor
- The provincial company arranges 4 experts from different vendors to verify, and the accuracy rate > 81%

Network automation, network intelligent simplification, artificial intelligence, machine learning and applications in 5G and other emerging fields will help improve user perception and satisfaction, accumulate experience for intelligent optimization and exploration in 5G and other emerging fields, and finally form artificial intelligence results in the network Optimize the scale application in the network service. Improve the overall strategy and service quality of China Mobile's network services through artificial intelligence application research, development and application of results in the field of network services.

6.3.4 Subscriber Complaint Handling

6.3.4.1 Background

With the network evolving, 2/3/4/5G network coexists and thus the problems of multi-RAT network, multiple elements and other issues make the subscriber complaint handling be completed. Traditionally, it mainly relies on the accumulated experience of experts, and requires high demands on operating labor. Subscriber complaint handling provides an end-to-end self-service solution for network management and operation, including complaint analysis and network fault handling. AI technologies are introduced to replace traditional methods and thus to improve the efficiency of fault solving.

This solution integrates four AI models and take the results of knowledge demarcation as the final output. Also, it extracts the core features that affect the results to support the demarcation. In order to facilitate the traceability analysis, it provides a visual display function for the whole process of demarcation.

At the late stage of complaint handling, the technology of abstract extraction is used to classify customer complaints automatically. It assists the human operator to analyze complaints, find potential problems in time and improve the accuracy of reply. As a result, repeated complaints can be greatly reduced. At the same time, it analyzes the complaint handling process and the results in the receipt, gives accurate and reliable reasons for complaints cascading to improve the efficiency of manual verification.

6.3.4.2 Solution Overview

Automatically dock with the EOMS, receive network tickets in real time, realize the automation of submission, delimitation and result confirmation to the EOMS. Also it can automatically return orders, dispatch orders or provide reference suggestions for complaints handlers. Through speech recognition, speaker segmentation, self-supervised learning, natural language processing and other AI technologies, the system can analyze the full number of complaints, extract the summary of complaints, and obtain the core information of complaints. After analyzing the complaint handling process and the results in the receipt, the accurate and reliable reason cascading is given to improve the efficiency.

- Automatic docking signaling side wireless side data to define seven categories including Wireless, Core Network, Communication Services, User Terminals, User Terminal Services, User Signing Services and No Exception.
- Combine curing experts to delimit the knowledge base.
- Deeply mine the correlation of alerts, use machine learning models to combine the periodic changes of network data with trend items, holidays and other influencing factors to fit the development trend of data for anomaly detection.
- Based on the time sequence characteristics of the signaling data to realize the comprehensive decision-making of the user's 24-hour fault situation.
- Considering the time and space factors to realize the accurate positioning of poor quality area.
- Abstract the complaint information and cluster the results for statistical analysis.
- Based on the knowledge graph and natural language processing technology to establish the alert expert knowledge base and provide the suggestions of alarm processing measures.
- Use natural language processing technology for receipt quality inspection instead of manual quality inspection personnel.

6.3.4.3 Application and Performance

At present, Network Self-service Robot system has been put into production. The complaint handling time is shortened from 90 hours to 41.21 hours, and the complaint location and demarcation time is shortened from 4 hours to 15 minutes. It is estimated that 45000 person days can be saved annually. According to the estimation of 700 RMB / person day, about 31.5 million RMB can be saved every year. The system assists operation and maintenance personnel to analyze complaints and reduce repeated complaints. Also it greatly improves the quality inspection efficiency.

6.4 Use Cases of Network and Service Optimization

6.4.1 NR Network UE Throughput Optimization

6.4.1.1 Background

Wireless network parameter configurations are subject to scenarios. There are thousands of parameters related to the air interface. Different parameters, such as handover, coverage, and power control parameters, have different impact scope on performance counters. The combination of parameters increases exponentially. It is difficult to achieve the optimal combination only through manual commissioning due to many types of parameter optimization, wide value ranges, complex scenario factors, and mutual dependencies between parameters. There are millions of parameter combinations.

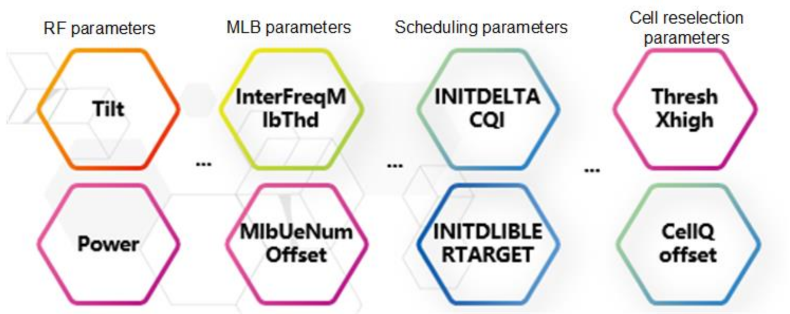

Figure 6-11 Thousands of parameters related to the air interface

In addition, wireless network scenarios are complex and diversified, and parameter settings need to vary depending on scenarios. Therefore, a large number of experts are required for analysis and processing, and it is difficult to achieve the optimal efficiency and performance. Traditionally, only expert experience can be used to analyze problems and optimize parameters. However, the efficiency of manual optimization on the entire network is low. In certain cases, the parameter settings used in a cell may bring negative gains in other cells.

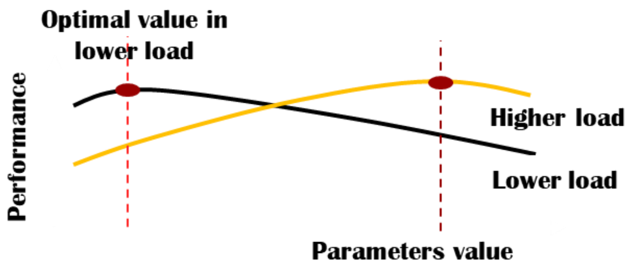

Figure 6-12 Parameter settings varies in different scenarios

6.4.1.2 Solution Overview

For RAN O&M, multi-parameter optimization is the most basic capability and all optimization tasks are performed based on parameter adjustment. The objective of the multi-parameter optimization solution is to ensure that all parameters of each cell can be automatically adjusted without affecting network KPIs. The vigorous development of AI technologies makes automatic multi-parameter optimization possible. With automatic multi-parameter optimization, the RAN Manager:

- Obtains the parameter optimization area and optimization objective, such as the target network KPI values, from the NMS.
- Obtain data. The RAN Manager automatically collects live network data (including MR data) based on optimization requirements and preprocesses the data, including data filtering and association.
- Automatically set scenario-specific parameters. The deep learning AI algorithm is introduced to the RAN manager to perform joint modeling analysis on KPI data of a large number of cells. In addition, the RAN manager automatically identifies networking scenarios based on the collected MR data on the live network and configures initial parameters based on the scenarios.

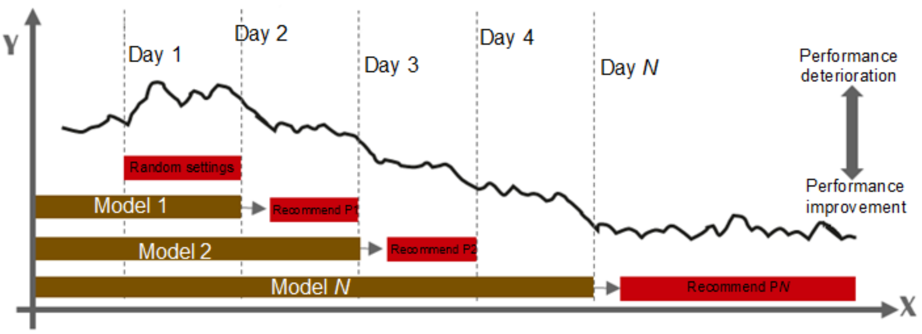

Figure 6-13 Iterative optimizing the model

- Performs automatic iterative optimization. The RAN manager uses live network data and machine learning AI algorithms to perform fast iterative optimization for multiple times and evaluate the impact of different parameter groups on network performance. This RAN Manager automatically configures optimal parameter combinations for cells based on different target settings to improve network performance (RRM parameters). In this way, network problems such as load balancing issues can be resolved.

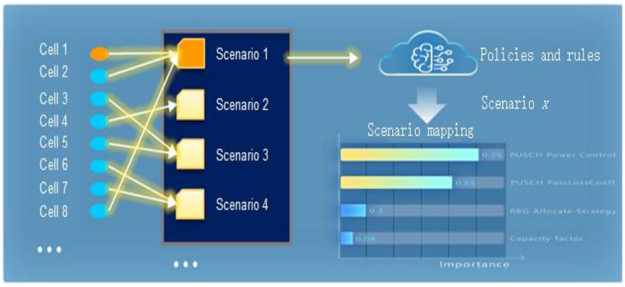

Figure 6-14 Mapping different scenario with proper parameters

6.4.1.3 Application and Performance

To cope with new network challenges, an operator in middle China has been exploring and innovating network intelligence and automation capabilities since 2019. In the traditional routine optimization process, telecom operators need to manually identify network coverage or capacity problems through DTs or network KPI statistics. Based on manual optimization experience, the RAN Manager provides advice on parameter adjustment, such as RF and network configuration, and then issues the network optimization policy. After the AI technology and automation capability are introduced, modeling is performed based on the multi-dimensional characteristics of cells on the live network, such as coverage, networking, traffic, and radio parameter configuration, and the average UE throughput in a cell. The RAN Manager can automatically identify low-rate areas on the network and automatically optimize 13 power parameters that are closely related to the single-user throughput of the cell based on this model while ensuring the overall network performance. In the Luoyang city, the average downlink UE throughput (more than 1000) cells increases by 14.5%.

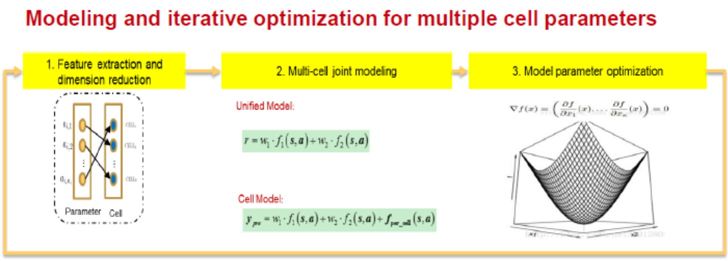

Figure 6-15 Modelling and iterative optimization for multiple cell parameters

In addition, the RAN Manager can automatically identify the cells subjected to load imbalance based on the live network data. Based on cell configurations and traffic models, the RAN Manager adjusts related parameters and load balancing policies to achieve the optimal balancing relationship between sectors. In the pilot area the load balance rate of the multi-band and multi-layer network increases by 75%, and the optimization efficiency is greatly improved.

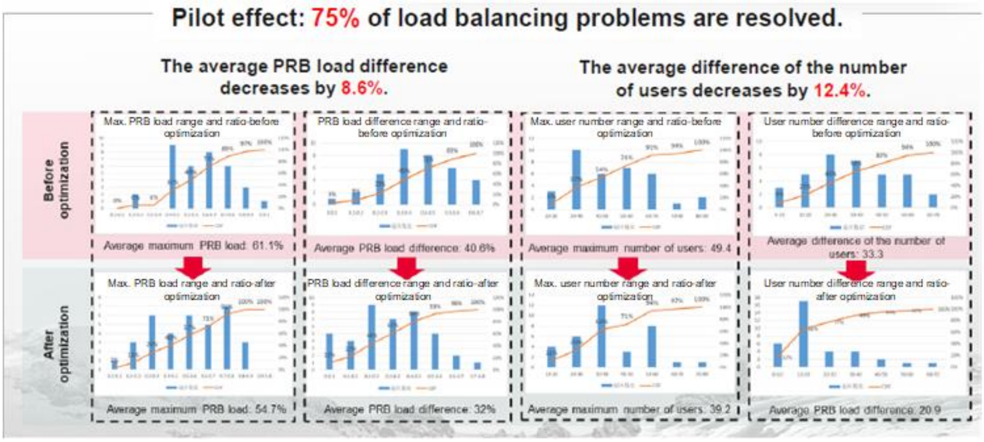

Figure 6-16 Pilot result of load balance with intelligent multiple parameter optimization

6.4.2 NR Network Coverage Optimization

6.4.2.1 Background

Massive MIMO is an evolution form of multiple-antenna technology, and is widely regarded as a key 5G network technology. This technology integrates more RF channels and antennas to implement three-dimensional precise beam forming and multi-stream multi-user multiplexing. Massive MIMO achieves better coverage and larger capacity than traditional technologies. In contrast with 4G massive MIMO that supports more than 200 broadcast beam combinations, 5G massive MIMO supports thousands of broadcast beam combinations. The pattern adjustment scope varies according to AAU types. Pure manual configuration and adjustment of broadcast beam combinations cannot achieve the optimal performance of massive MIMO due to its complexity. When massive MIMO modules are deployed on a large scale, the adjustment workload is heavy, and it is difficult to complete the adjustment manually.

According to the test results of multiple operators on the live network, massive MIMO intelligent optimization can improve the RSRP and UE throughput and maximize operators' ROI.

6.4.2.2 Solution Overview

When massive MIMO intelligent optimization is enabled, the RAN Manager:

- Obtains coverage optimization areas and objectives, such as the proportion of weak coverage areas, from the NMS.
- Obtains DT data, performance counters, traffic statistics, engineering parameters, configuration parameters, and other basic information, including electronic maps, antenna patterns, frequency bands, and AAU types.

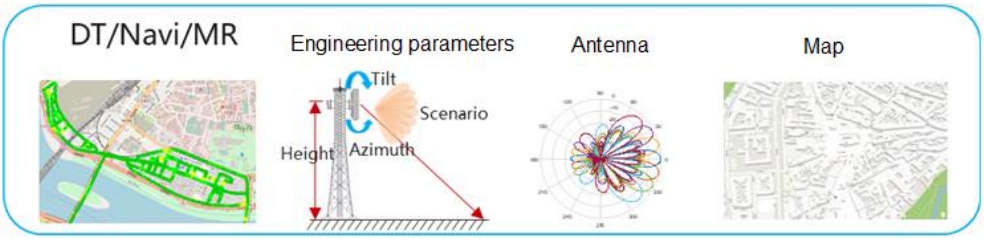

Figure 6-17 Multiple dimension of data collection

- Creates grids for DT/MR data, identifies problematic grids, and converges them into problematic areas. Then, the RAN Manager selects the best scenario-based beam, azimuth, and down tilt configurations for problematic cells. In this step, antenna hardware must meet the corresponding configuration requirements.

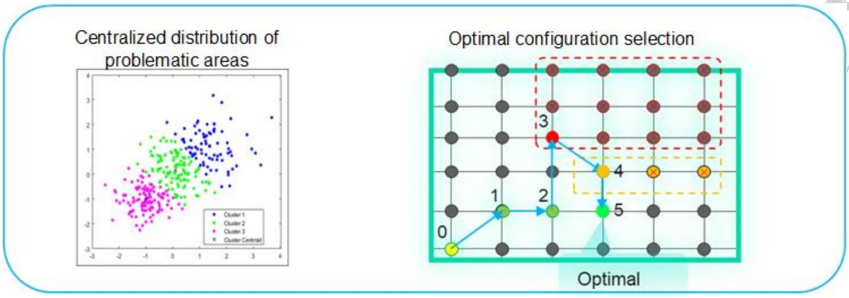

Figure 6-18 Locate the problematic area and find the optimal configuration

- Performs iterative reinforcement AI learning based on the preset optimization objectives to obtain the optimal optimization advice.
- Automatically delivers the massive MIMO pattern parameter combination, down tilt, and azimuth parameters of problematic cells and their neighboring cells base on Massive MIMO pattern common AI model.
- Evaluates and verifies the optimization advice based on user experience after issuing the optimization advice. If the KPIs do not meet the target requirements, the RAN manager rolls back the optimization advice.

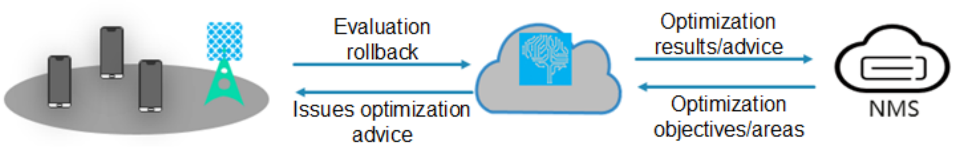

Figure 6-19 Evaluate and verify the optimization advice

6.4.2.3 Application and Performance

In a typical operator application scenario, the RAN Manager interconnects with the NMS through an open API. The NMS delivers the network coverage optimization objectives and areas to be optimized to the RAN Manager. The RAN manger sends the final optimization result and optimization advice of each round to the NMS of the operator.

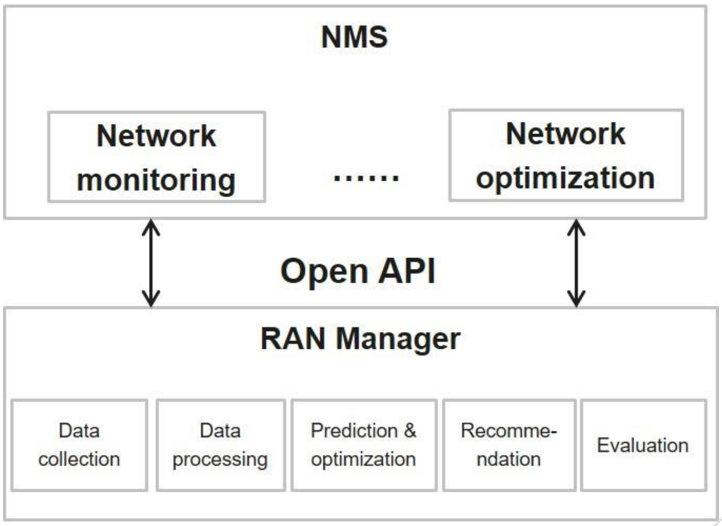

Figure 6-20 Network framework in a typical operator application scenario

In 5G network deployment scenarios, the number of 5G UEs is small. One operator in middle China applies AI technologies and automation to optimize massive MIMO broadcast beams. The RAN Manager can automatically identify problems found during drive tests, such as weak coverage, poor SINR, overlapping coverage, overshoot coverage, and frequent handovers. Based on experience rules, coverage prediction, and 5G weight parameter optimization, the RAN Manager provides parameter adjustment advice on mechanical tilts, azimuths, and broadcast beam weight. In this way, 5G coverage and performance can be quickly improved to ensure better experience.

This solution increases the average coverage of 5G massive MIMO cells by 15.8% and the road coverage by 91%. In addition, the optimization efficiency is significantly improved in contrast with traditional optimization methods.

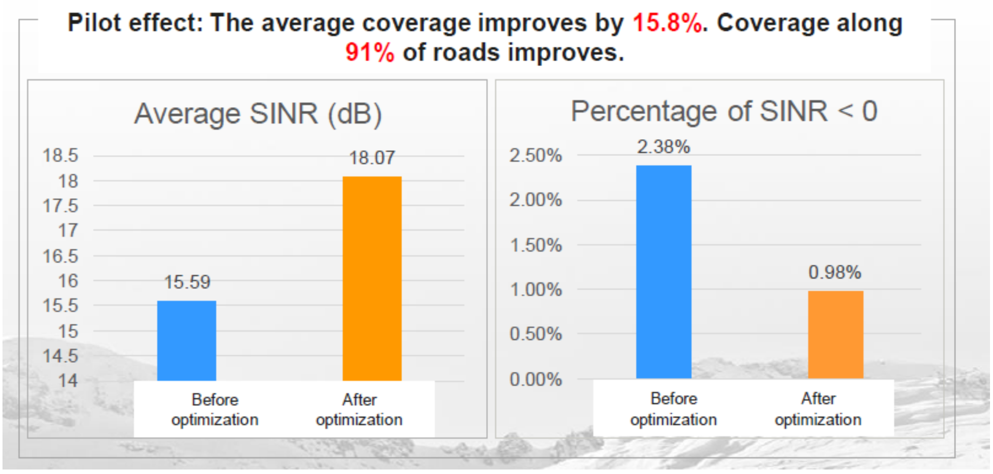

Figure 6-21 Pilot result of NR network coverage optimization

6.4.3 ML-Based MU-MIMO Scheduler

6.4.3.1 Background

Massive MIMO with massive number of antennas is one of the key enhancements of 5G. Narrow beams can be used in the regions of high user density whereas wider beams could be used in the regions of low user density.

It brings opportunities to further enhance cellar network, in terms of spectral efficiency and user throughput. By utilizing the radio resources more efficiently, the next generation 5G promises to bring much better services to consumers, to open more business opportunities and revenues to operators.

However, it comes along with a great potential challenge. Massive MIMO means there are large number of beams and user layers needed to be managed. In order to realize the full potential of multi-user massive MIMO, the scheduler is needed to solve very complicated problems, which include to figure out the best set of beams. Simply it figures out that combinatorial complexity of selecting 4 beams from 32 is more than 100,000 choices. The challenge here is to be able to design advanced scheduler that can optimize the spectral efficacy within practical compute complexity. Based on the channel sensing measurements, an optimal beam-former configuration might be derived, which can improve the user throughput.

6.4.3.2 Solution Overview

The basic idea on how to model this problem is to decompose the multi-step selection problem, without using programming, into simple sub-problems in a recursive manner.

For every TTI, the objective is to find the set of beams that maximize the Q value, in this case, is max {Sum UE-PF}. (UE Proportional fair), which is key system performance indicator, also commonly used to evaluated scheduler performance.

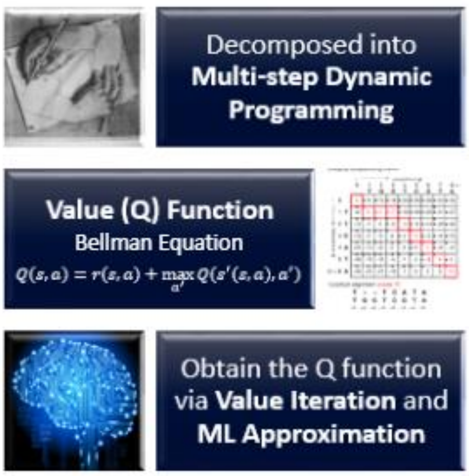

Figure 6-22 Deep Q-learning or reinforcement learning

Then the Q value function is the state as the function pair. The action in simply, which is the beam selection. The state is designed to capture the beam UE_PF as the function of its channel condition and into beam interference. And the reward is the net benefit, i.e. beam on the sum UE_PF when adding a new beam UE. It is important here to use MU_PF rather than SU_PF, which means multi-user proportional fair. To use MU_PF here is to take the inter-beam inference into account. Due to the recursive property of such a dynamic programming, the value function can be represented as immediate reward as future reward possible which is called Bellman Equation. Using this property, Q Value can be obtained via value iteration. However, such value iteration is hard to store conventionally as the state space is huge. So rather a deep neural network is used to approximate the value function here. Such model free dynamic program is also called Deep Q-Learning or Reinforcement learning.

6.4.3.3 Application and Performance

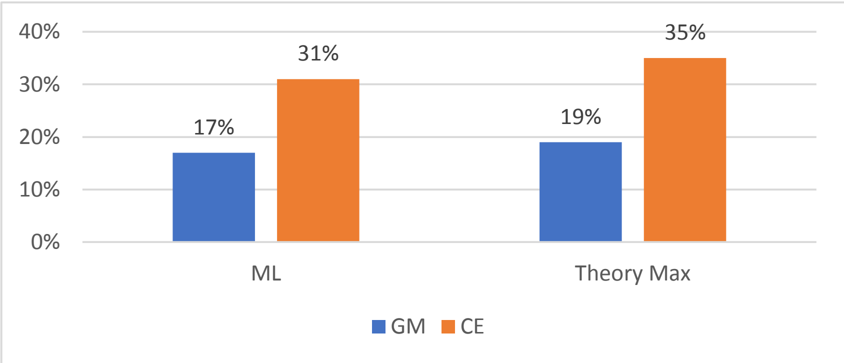

Figure 6-23 Simulation results of ML-based MU-MIMO scheduler

This diagram shows the simulation evaluation of ML algorithm, compared with the traditional greedy algorithm in the current existing product and the best theoretical possible in this case. The theoretical max can be obtained via exhausted search.

The highlight the result is:

- The ML scheduler (algorithm) can achieve close to theoretical max.
- The ML scheduler (algorithm) can achieve very good performance gains 17% in geomean (GM) UE throughput and 31% in cell edge (CE) user Tput comparing to traditional algorithm.

6.4.4 Link Adaptation

6.4.4.1 Background

5G spectrum is limited and frequency efficiency is very important for operators. In 5G network, DL spectral efficiency and throughput may be affected by inter-cell interference. The current DL Link Adaptation is optimized towards slowly varying and stationary channel variations. The network may show suboptimal performance when adjusting to interference created by burst traffic, such as low spectral efficiency, low throughput. It will impact the end user experience.

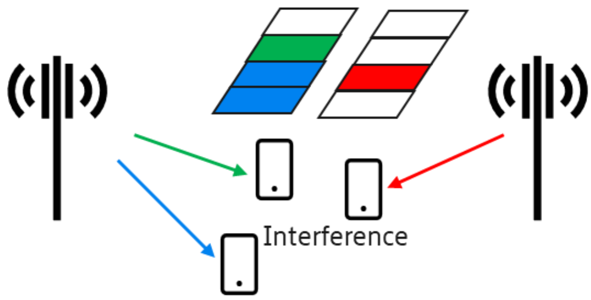

Figure 6-24 Interference created by burst traffic

It is a big challenge how network adapt the interference to get better performance. New technology AI can be used into network to recognize burst interference and optimize link adaptation.

6.4.4.2 Solution Overview

Using neural network on RAN, introduces a Machine Learning algorithm for steering of the existing Link Adaptation. The ML algorithm is trained to recognize refined interference scenarios based on the history of the neighbor cell activities and UE signal quality.

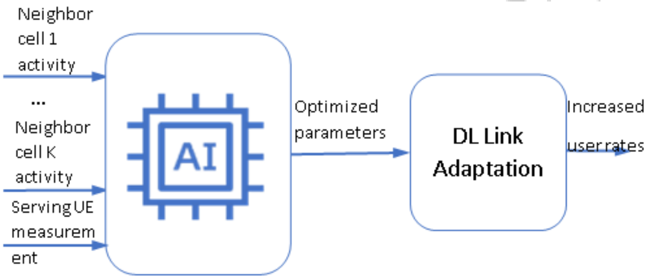

Figure 6-25 Intelligent DL Link Adaptation

It can optimize radio link performance using pattern recognition for each radio link. Collect adjacent cell data in real time, use that together with mobile in order to make smart link adaptation to better fit air quality. Proper tuning the link adaptation by a Machine Learning algorithm is built on the history of neighbor cell activity and serving UE measurement.

The ML algorithm tune the link adaptation dynamically in time and individually for each UE for a short period of time (sub-seconds). This replaces the module based constant homogenous parameters setting.

6.4.4.3 Application and Performance

Intelligent link adaptation can improve end user performance and spectrum efficiency. It will be no UE dependency and can get more optimistic scheduling which can allow user to schedule higher data rate.

The biggest gain can get from overlap cells outdoor. For indoor scenario, only single cell and no overlap cells and it will not have any gain since intelligent link adaptation takes into consider other cell interference. If there are no other cells around the cell, there is no gain. There will have big value for multi cell in city center and urban area with intelligent link adaptation.

Based on simulation result, the cell edge downlink throughput can be up to 50% gains.

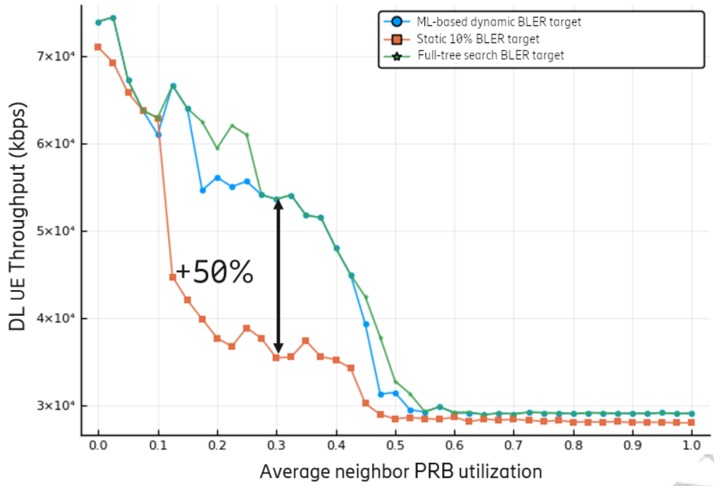

Figure 6-26 Simulation result of intelligent link adaption

With simulations, it shows that in areas with high cell overlap and medium to high loaded, the spectral efficiency can be improved up to 15%.

The gain is present in light to high load, not overload sites. ~70% cell has <50% PRB load. This function can be used in ~70% cell.

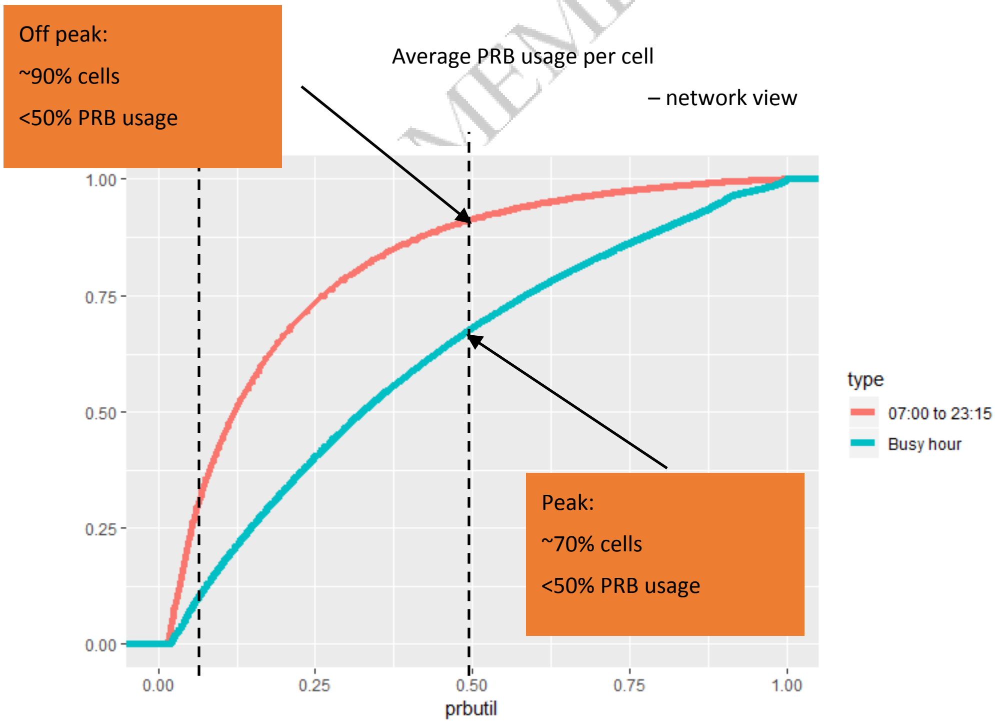

Figure 6-27 Simulation result of intelligent link adaption

According with 4G test result, the cell edge downlink throughput can be up to 50% gains. In areas with high cell overlap and medium to high loaded, the spectral efficiency can be improved up to 15%. Considering 5G scenario, the gain should be the same as 4G.

6.4.5 Load Balancing Based on the Virtual Grid Technology

6.4.5.1 Background

With the continuous development of communication technologies, brings more subscribers with higher data consumption. As a result, the cell load surges and the load among multi-carrier cells is unbalanced. Therefore, the load distribution among cells needs to be adjusted based on the load balancing policy to improve user experience.

In the current balancing policy, the selection of the balancing UE and the balancing target cell is based on capabilities and some cell-level information. Relatively blindly, the load of the target cell may be low. However, the selected UE is located in the area with poor or no coverage of the target cell, resulting in invalid measurement or blind handover failures.

By introducing virtual grids, you can predict the radio characteristics of the neighbor cells where UEs are located, and rapidly and accurately implement load balancing between cells on the multi-frequency layer, improving resource utilization and user experience.

A virtual grid is a space division method based on geographical location information to obtain the signals of multiple intra-frequency cells under the current environment of a UE, and then divides areas based on the cell and signal quality. By collecting statistics on the radio features of each grid (such as inter-frequency adjacent cell coverage), we can deploy more streamline network strategies.

6.4.5.2 Solution Overview

The virtual grid-based balancing solution includes the following aspects:

- Grid library building:
  Based on the historical measurement reports and handover information of UEs, the AI algorithm is used to construct virtual grids and obtain the relationship between the grid-level UE and the wireless coverage of surrounding cells.
- Grid database update and evaluation:
  - Writes UE information into virtual grids and updates them in accordance with the real-time intra-frequency measurement information of UEs.
  - This feature collects statistics of real-time intra-frequency and inter-frequency measurement information and handover information of UEs, evaluates neighbor cell information in the grid database, and updates neighbor cell information as required.
- Virtual Grid Application:
  The system monitors the load of each cell in real time. If the load unbalance conditions between cells are met, the system deduces the UEs that can be balanced and their target cells based on the cell load, cell characteristics, UE characteristics, and relations between UEs and surrounding cells (virtual grid information), and provides execution suggestions.

  While executing load balancing based on virtual grids, perform availability assessment on virtual grids. When availability is low, initiate a request for updating virtual grids.

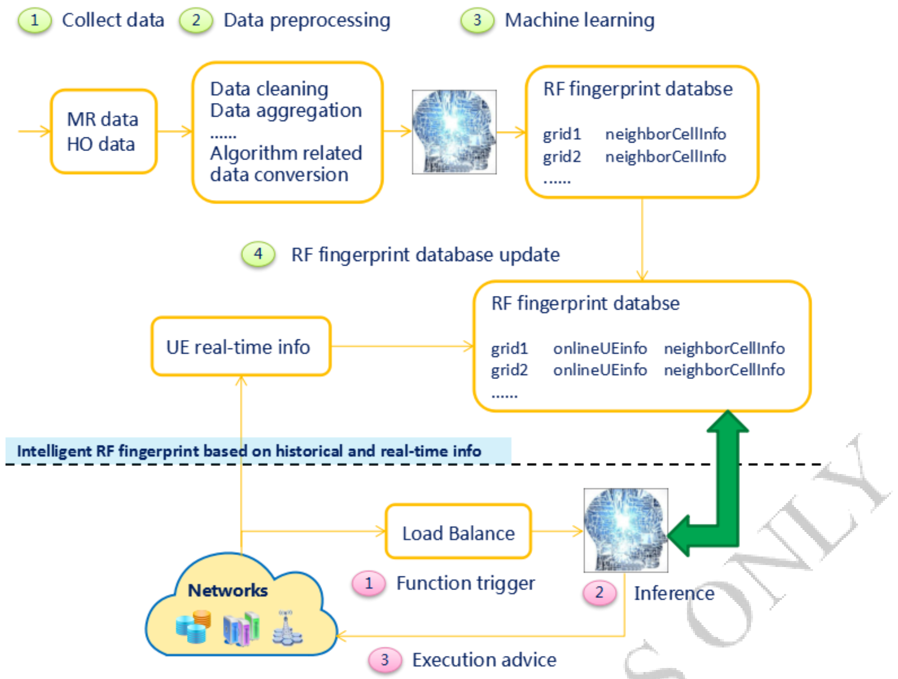

Figure 6-28 Basic flow chart

6.4.5.3 Application and Performance

Application scenario: Load balancing in a multi-frequency network architecture.

Expected application effect:

In a traditional load balancing cell, the same policy is applied to almost all UEs, but the coverage and performance of neighbor cells in different areas are different. That is, the target cells where UEs can be balanced may be different.

The grid information can be used to determine whether a UE is suitable for balance and the target cell that can be balanced, thus improving the overall performance.

- No balance cell is available for the grid where the UE is located: The UE does not perform balance to reduce handover failures.
- The grid where the UE is located has cells that can be balanced: The UE performs blind balancing to reduce inter-frequency measurement, or performs targeted measurement only for the cells that can be balanced to improve the measurement efficiency, improve the UE migration speed, shorten the load unbalance time, improve the resource usage, and improve the overall throughput of the region.

Table 6-1 The actual application effect in China Mobile QuanZhou (23 cells of 8 eNBs for a week)

Parameter	Normal LB	RF Fingerprint LB (non-measurement)	Improvement
High Load Time of Cell (s)	1542918	1338440	13.25%
Number of Cell Load Balance Occurred	93853	79339	15.46%
Number of Load Balance intra-system Measurement Report	1022960	169424	83.44%
Number of Load Balance intra-system Measurement Configuration UE(Due to Enhanced Load Balance)	909375	415774	54.28%
Handover success rate based on load balance	stable		
Basic KPI (such as RRC Establishment Success Rate, E-RAB Setup Success Rate, E-RAB Drop Rate)	stable		

The result shows that high load time, load balance times, MR reports and measurement configuration are all decreased after applying RF Fingerprint LB(non-measurement), which indicate that some UEs can move to the other cells more easily without measurement, therefore load balancing efficiency is improved, and some inter-frequency measurements are saved.

Note: since cell load is not high enough, the improvement of throughput is not observed.

6.4.6 QoE Optimization

6.4.6.1 Background

With the deployment of 5g network, many 5g native applications, such as cloud VR, 8K video are booming. Cloud VR service needs high transmission bandwidth and is sensitive to delay. Compared with the traditional audio and video services, the user experience (QoE) of cloud VR is more vulnerable to the fluctuations of wireless transmission, resulting in stuck, mosaic and vertigo. Traditional semi-static QoS framework can't efficiently satisfy diversified QoE requirements of different applications. At the same time, QoE estimation through user's interactive information in the application server usually results in a large delay, can't prevent the decline of user experience.

The "QoE Optimization" use case usually involves a network element function of gNB and Local server/MEC which collects service requirements of the VR application from MEC/local server and radio status of UEs the BTS. Then, using data analytics and ML inference, predict the UE's radio status e.g bandwidth in next 10 millisecond, and coordinate VR streaming encoding rate and radio resource scheduled to optimize the user experience and prevent QoE degradation (video stream jitter, mosaic etc.) as the fluctuations of wireless transmission. It is expected that QoE optimization via the intelligent collaboration between the application server and RAN can help deal with wireless transmission uncertainty and improve the efficiency of radio resources, and eventually improve user experience.

6.4.6.2 Solution Overview

As shown in Figure 6-29, an example of intelligent management function is introduced. The function of intelligent management function is to deploy and manage intelligent applications and provide the data and management channels to application server and BTS. Through Intelligent management function, three artificial intelligence application modules are deployed for QoE optimization: AI based Application recognition, AI based QoE evaluation and AI based wireless bandwidth prediction. AI based Application recognition module is to identify VR applications by using the transmission pattern via ML, AI based QoE evaluation is to evaluate the score of user experience of the VR applications according the jitter, mosaic etc… which can be recognized by the traffic pattern by AI, AI based wireless bandwidth prediction is to forecast radio channel quality in next 10-20ms. A close loop QoE optimization is realized via interaction among 3 AI modules. .

Using QoE optimization procedure of a VR application as example, firstly, the application recognition module identifies the VR application among the number of broadband applications which are transmitted by the BTS. The QoE evaluation module monitors the VR applications user experience online. When the user's signal-to-noise ratio becomes worse, the wireless bandwidth predicts that the transmission rate will decline, and the QoE starts to deteriorate, it can inform/suggest BTS and application server to act and prevent QoE decline. e.g. the VR application server is informed/suggested to reduce the coding rate and keep the stream smooth, and BTS is informed/suggested a min reserved PRB which satisfy the required bandwidth for the service with lower coding rate. Since the action is taken according to QoE predication, the corresponding code rate and schedule rule adjustment have been completed before the wireless bandwidth changes. When the fluctuations of wireless transmission happen, the user's experience will be affected little.

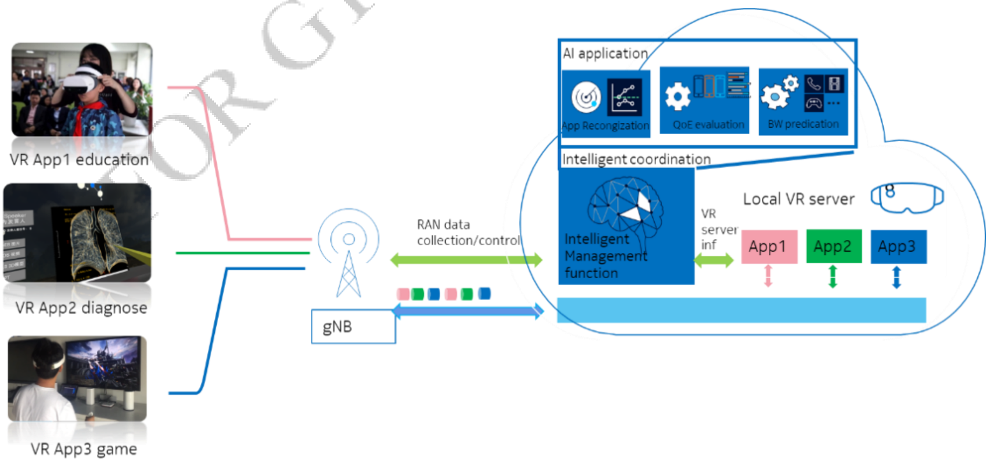

Figure 6-29 Intelligent management function used in VR APP for QoE optimization (example)

6.4.6.3 Application and Performance

In the case trial, the VR application server and intelligent management function were introduced, which is located in an aggregation transport equipment room, where is approximately 2 km from the 5G BTS. An example of VR cloud gaming over 5G network were used to test user experience score with/without QoE optimization. As the test results, the estimation accuracy of Application recognition evaluation Wireless bandwidth prediction is more than 90%. Network adjustment latency, which is the VR encoding rate adjustment latency after radio quality change, reduce from 20s to 1s, as action can be taken earlier via wireless bandwidth predication. User QoE Score increase from 40% (bad user experience) to 90%, which means a fluently VR stream.

Where user-specific 3D game video is rendered in the Edge App server:

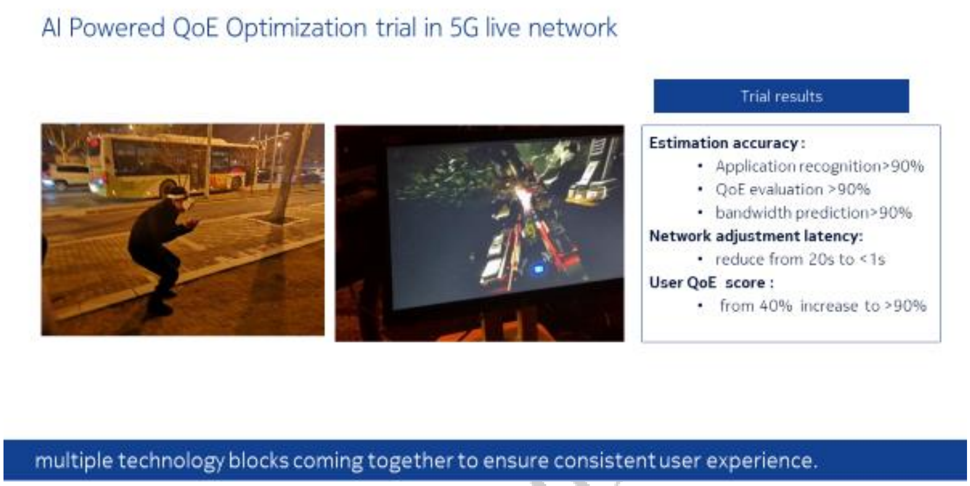

Figure 6-30 AI powered Cloud VR experience trial on China Mobile live 5G network

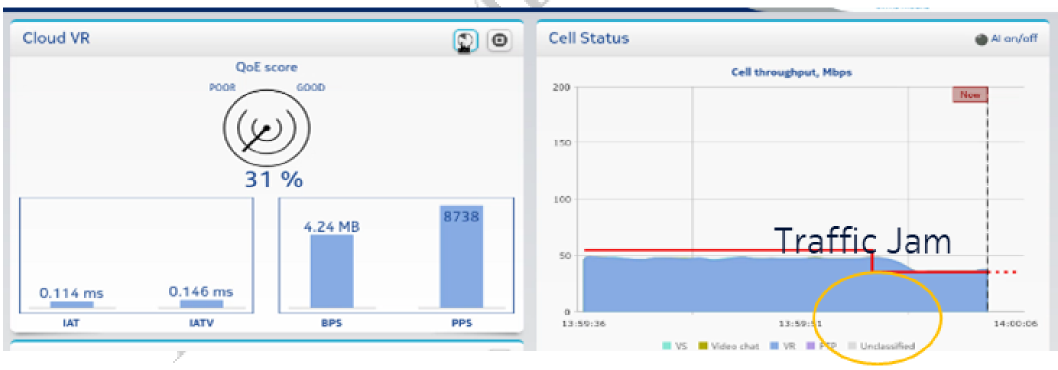

Figure 6-31 trial results: VR game QoE evaluation without QoE optimization

VR game QoE score detected by AI reduce to 31%, which indicate low QoE, when congestion happens.

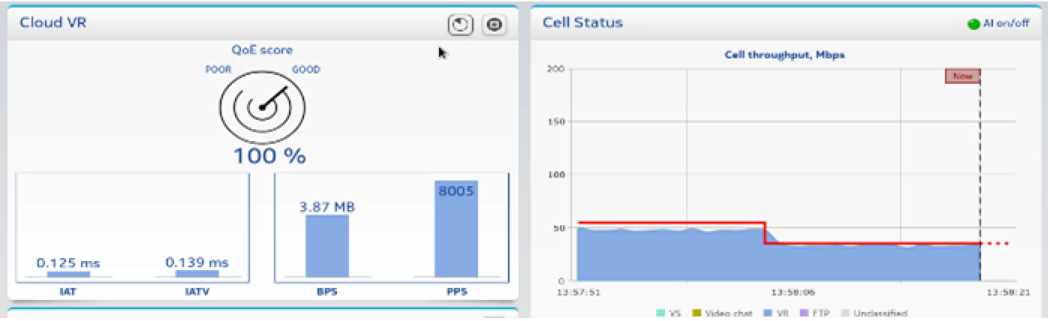

Figure 6-32 trial results: VR game QoE score with QoE optimization

VR QoE score detected by AI reduce increase to 100%, which indicate high QoE, by AI command VR server decrease the encoding rate. VR QoE increased in short time (<0.1s), and User didn't notice the VR QoE change.

The intelligent management function is deployed to manage intelligent applications and provide the data and management channels to application server and BTS.

AI/ML-based QoE optimization via the intelligent collaboration between the application server and RAN improve the user experience. The 5G trial results shown with AI/ML-based QoE optimization, network adjustment latency for the fluctuations of wireless transmission reduce 10 times, reduce from 20s to 1s, as action can be taken earlier via wireless bandwidth predication. And User QoE Score increase 3 times, from 40% (bad user experience) to 90% (VR stream play fluently).

6.4.7 Edge QoS

6.4.7.1 Background

In the 5G era, more than ever before, the access network forms and contents of services are enriched. With its position in the network, MEC makes it possible to provide near-real-time services for latency-sensitive, user-sensitive, and service-sensitive applications. In the campus application scenario, the MEC can obtain the service requirements of the application and the UEs connected to the application service from the app. Therefore, a QoS control policy can be delivered to the base station in accordance with the service requirements of the app, and the service QoS of some UEs can be adjusted to meet the service requirements of the app.

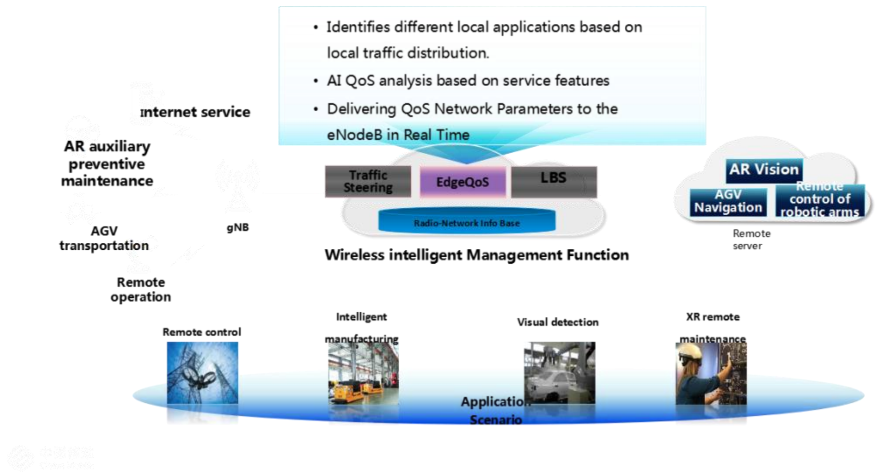

Figure 6-33 Edge Qos application scenario

6.4.7.2 Solution Overview

The wireless network information, UE context information and measurement information reported will provide the location perception and wireless environment perception for the Apps on the MEC. Meanwhile, in the UPF/MEC co-deployment scenario, the MEC can obtain the UE ID to provide the UE ID perception for the Apps. Apps obtains the near real-time perceived data, generates the near real-time control policy according to the application requirements and radio capability optimization algorithm, and then delivers it to the BTS for executing the scheduling policy action.

The gNodeB needs to abstract a service model in accordance with the function type that provides wireless network information. A function is mapped to a service model. When establishing a connection, the gNodeB notifies the MEC of the functions it supports, and the MEC determines which functions need to subscribe to RAN wireless network information. The MEC generates a control policy according to the service identification, application requirements and algorithm optimization. It selects some UEs or a certain QoS Flow under a UE's PDU Session, delivers the control policy to the BTS, and the BTS executes the QoS Flow-based scheduling optimization.

6.4.7.3 Application and Performance

In the uplink direction, the gNodeB needs to report the radio network information, UE context information, and measurement information to the MEC. In the downlink direction, the MEC delivers the service control policy and QoS parameters to the gNodeB in accordance with the application requirements.

In this case, the AI technology is used for service model identification and scheduling model optimization output. For example, based on various sensing data, the ML/DL algorithm is used to output a best scheduling model and deliver it to the base station for QoS guarantee.

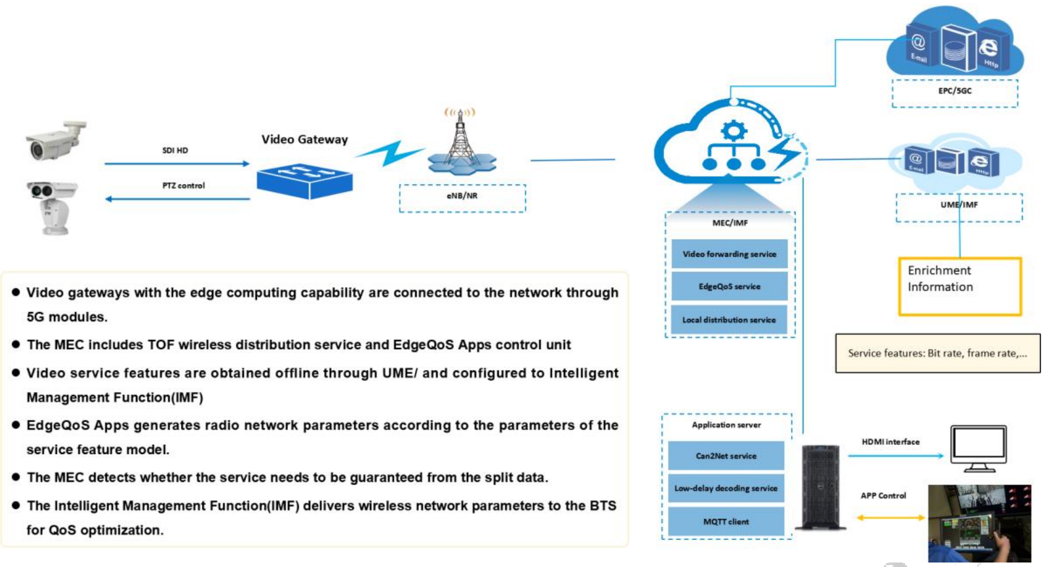

Figure 6-34 Edge Qos remote control guarantee application

6.4.8 Transport Network Optimization

6.4.8.1 Background

Nodes of transport network can be logically divided into three types: Access Layer Node, Convergent Layer Node and Backbone Layer Node. Access Layer Nodes are generally deployed as user access points. Convergent Layer Nodes are usually deployed at sites with convenient optical routing and a large number of optical cables. Multiple Access Layer Nodes converge at the Convergent Layer Node, and multiple Convergent Layer Nodes converge at the Backbone Layer Node.

The topology of the traditional transport network is chain architecture, which is easy for extension. However, this topology will split transport network into multiple isolated areas, which causes seriously affects on the capacity and performance of transport network when a single node or link fails. In view of this disadvantage, current transport network transforms into ring architecture. This type of transport network has high robustness and reliability where failure of single node or link will not lead to network partitions.

Based on the characteristics of ring architecture, in the traditional capacity management approach, the virtual network function (VNF) among the transport ring will be expanded to ensure ring capacity requirement of business transport, when the capacity usage rate of transport ring reaches a certain threshold. However, this management mode lacks information linkage between transport rings, which means different transport rings cannot perceive the service load among each other. This mode will lead the redundant expansion operation, resulting in low utilization rate of the overall capacity of the transport network, and increasing the construction cost. In order to utilize the network resources more efficient, operators put forward related manual-decision optimization plans. However, the formulation of the final optimization plan is highly dependent on the manual analysis and decision based on expert experience in related fields, leading the process tedious and time-consuming.

With the advent of 5G and saturation of telecom industry market, higher requirements are put forward for the guarantee of transport network capacity, that the guarantee of capacity is bound to be carried out with lower cost and higher efficiency.

6.4.8.2 Solution Overview

In order to ensure the optimum capacity utilization of transport network and avoid unnecessary capacity expansion operations, traditional optimization methods include:

- Keep the link connection between VNFs unchanged, and expand the capacity of targeted VNF(s) among transport ring with excessive capacity usage rate.
- Keep the link connection between VNFs and the capacity of each VNF unchanged, and add VNFs to form new transport ring.
- Keep the overall topology of transport network and capacity of each VNF unchanged, and adjust the service load on each transport ring.

However, when the capacity usage rate of a ring exceeds the pre-determined health threshold, the other rings are usually in a state of idle or low usage rate. Due to the lack of information linkage between transport rings, above methods cannot accomplish intelligent optimization on the transport network. Therefore, in order to maximize the optimization effect and reduce the labor cost as much as possible, the following method is preferred:

- Keep the capacity of each VNF unchanged, and then adjust the link connection between VNFs to achieve the globally optimum utilization. If the capacity usage rate of the transport ring is still over specific threshold, then carry out capacity expansion operation on the relevant VNFs among this transport ring.
- This method adjusts the network topology to improve capacity utilization and service load-balance effect. After topology optimization, if the capacity of transport ring is still unable to meet the service requirement, then the relevant VNFs can be expanded.

In order to achieve intelligent optimization method, Transport Network Optimization system is established, which consists of following stages:

- Stage 1: Data Processing. In this stage, the detailed information of transport network (e.g. network topology, coordinates of VNF, VNF capacity, peak-hour data flow on transmission ring, etc.) are collected. Moreover, parameters used for optimization policy management will be also input. These data are transformed into a common format based on normalization algorithms and will be further processed by following stages.
- Stage 2: Data Analysis. In this stage, based on the existed network topology and pre-processed data, Transport Network Optimization system analyses the relation amongst capacity of transmission ring, topology of transmission ring and the capacity of VNF. Besides, this system will calculate the global utilization rate of capacity.
- Stage 3: Decision and Output. In this stage, according to the pre-determined optimization policy, Transport Network Optimization system will decide the optimization plan, including optimized network topology and VNF capacity requirement among each transmission ring, based on the result of Data Analysis and AI/ML algorithm. This plan will be used by the operator to optimize the transport network reaching globally optimum capacity utilization

rate.

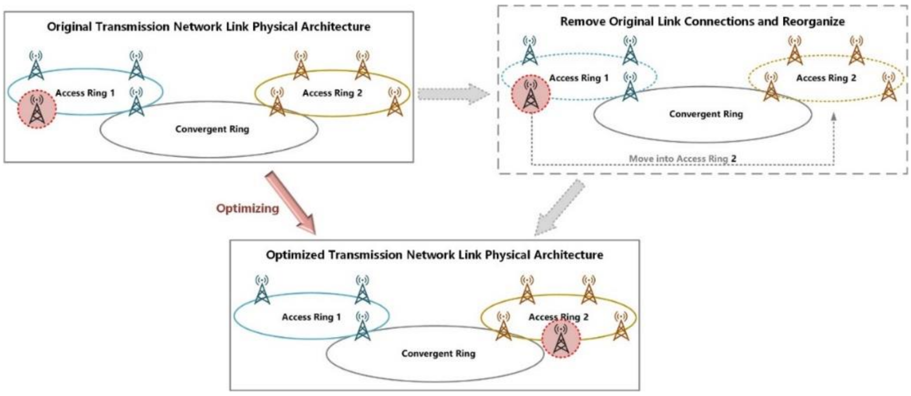

Figure 6-35 Original and Optimized Transport Network Link Physical Architecture

6.4.8.3 Application and Performance

Nowadays, Transport Network Optimization system has been put into production and provides the following effects:

- The time taken to formulate the final optimization plan has been shortened from 48 hours to 2 hours with approximate 1000 nodes;
- The capacity expansion cost has been reduced from 50-million RMB per year to 40-million RMB per year by applying the intelligent optimization system;
- The labor cost has decreased to 960 person-hour, saving 1500 person-hour compared to the traditional approach.

Transport Network Optimization system also assists operation and maintenance personnel to monitor, analyze transport network service-load and formulate the optimization plan, greatly improving the network performance.

7 Intelligent Network Level

7.1 Introduction

Intelligent network level describes the level of application of intelligence capabilities in the network management and control workflow. The participation of the human and telecom system in the network management and control workflow are important factors to evaluate the intelligent network level. For each intelligent network level, which tasks can be performed by telecom system, which tasks can be performed by human, and which tasks can be performed by cooperation of human and telecom system needs to be clarified. For example, in the highest intelligent network level, all tasks are performed by telecom system.

The industry benefit from common method for evaluating intelligent network level, which provides evaluation basis for measuring the level of an intelligent network along with its components and workflows, reference for gaps and priorities analysis for standardization works on intelligent network level and guidance to operators, vendors and other participants of telecommunications industry for roadmap planning.

7.1.1 Framework Approach for Classification of Intelligent Network Levels

According to the potential categorization of the tasks in a general network management and control workflow (including intent management, collection, analysis, decision and execution), a framework approach for classification of intelligent network level is introduced as following, which is used for evaluating the intelligence capabilities of telecom system.

Note: the following framework approach for classification of intelligent network level is based on the framework approach for classification of autonomous network level defined in 3GPP TR 28.810[7]. And the framework approach for classification of mobile network management and operation intelligence levels defined CCSA TC7 [34].

Table 7-1 Framework approach for classification of autonomous network level

Network autonomy level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent management
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary intelligent network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate intelligent network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.						

- **Level 0 manual operating network:** No categorization of the tasks is accomplished by telecom system itself.
- **Level 1 assisted operating network:** A part of the execution and awareness tasks are accomplished automatically by telecom system itself based on human defined rules. At this level, telecom system can assist human to improve the execution and awareness efficiency.
- **Level 2 preliminary intelligent network:** All the execution tasks are accomplished automatically by telecom system itself. A part of the awareness and analysis tasks are accomplished automatically by telecom system itself based on human defined policies. At this level, telecom system can assist human to achieve the closed loop based on human defined policies.
- **Level 3 intermediate intelligent network:** All the execution and awareness tasks are accomplished automatically by telecom system itself. A part of the analysis and decision tasks are accomplished automatically by telecom system itself based on human defined policies. At this level, the telecom system can achieve the closed loop automation based on the human defined closed loop automation policies.
- **Level 4 advanced intelligent network:** All the execution, awareness, analysis and decision tasks are accomplished automatically by telecom system itself. And intent management tasks can be partly accomplished automatically by telecom system itself based on human defined intent translation policies. At this level, telecom system can achieve the intent driven closed loop automation based on human defined intent management policies, which means the telecom system can translate the intent to the detailed closed loop automation policies and evaluate intent fulfillment information (e.g. the intent is satisfied or not) based

on human defined intent management policies.

- **Level 5 fully intelligent network:** The entire intelligent network management and control workflow is accomplished automatically by telecom system without human intervention. At this level, telecom system can achieve the whole entire intelligent network management and control workflow to satisfy the received intent.

Note: Above framework approach for classification of intelligent network level are applicable for evaluating the intelligent network level from both applicable scope (including NE, domain, cross domain) and applicable scenario perspective. The overall intelligent network level of the whole telecom system is a comprehensive reflection of intelligent network level of the individual applicable scope and applicable scenarios, which means in fully intelligent network level, the telecom system can achieve the whole intelligent for all applicable scopes and applicable scenarios.

7.2 INL Evaluation of Typical Use Cases

7.2.1 Energy Saving

Based on the framework approach for classification of intelligent network level, the use case of online energy saving wireless networks defined in clause 6.3.1 achieves intelligent capabilities of level 3:

- Intent management (Human): Performed by human
- Awareness (Telecom System): The RAN Manager can automatically collect network data and generate traffic distribution information based on the collection results.
- Analysis (Human & Telecom System): Based on the manually specified areas and energy saving objectives, the RAN Manager can automatically generate energy saving solutions specific for network scenarios and traffic conditions.
- Decision (Human & Telecom System): After the RAN Manager generates an energy saving solution, the network automatically evaluates the energy saving solution, continuously performs iterative optimization, and finally provides the optimal solution.
- Execution (Telecom System): The RAN Manager automatically delivers the energy saving solution and automatically configures the entire network and sites.

Table 7-2 Level evaluation of Energy Saving

Network autonomy level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent management
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary intelligent network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate intelligent network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.						

To evolve to a level 4 intelligent network, the RAN Manager needs to provide the intelligent capability of predicting network traffic and generating energy saving policies automatically based on capabilities of level 3.

7.2.2 NR Network Coverage Optimization

Based on the framework approach for classification of intelligent network level, the use case of NR network coverage optimization defined in clause 6.6.4 achieves intelligent capabilities of level 3:

- Intent management (Human): performed by human
- Awareness (Telecom System): The RAN Manager can automatically collect coverage-related performance data and generate coverage geographic distribution information (for example, RSRP distribution information based on geographic grids) based on the collection results.
- Analysis (Human & Telecom System): The RAN Manager can automatically analyze and determine the root causes of coverage problems, such as identifying top N abnormal cells, based on the specified coverage problem analysis policies. The RAN Manager generates a coverage optimization and adjustment solution (such as the optimal massive MIMO coverage scenario, azimuth, and tilts of top N abnormal cells and their neighboring cells) based on the manually specified coverage optimization and adjustment policies, such as range of the massive MIMO pattern, azimuth, and tilts.
- Decision (Human & Telecom System): For an automatically generated optimization solution,

the RAN Manager can evaluate the coverage optimization solution and its impact on network performance based on the manually specified coverage optimization and adjustment policy. Finally, the RAN Manger automatically determines the coverage optimization and adjustment solution to be executed based on the manually specified coverage optimization and adjustment policies.

- Execution (Telecom System): Based on the automatically determined optimization and adjustment solution, the RAN Manager automatically adjusts related network parameters.

Table 7-3 Level evaluation of NR network coverage optimization

Network autonomy level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent management
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary intelligent network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate intelligent network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.						

Level-3 policy driven closed-loop coverage optimization automation is achieved in the use case. As the next-step evolution of intelligent networks, intent driven closed-loop coverage optimization automation based on specific service assurance intents for certain scenarios needs to be implemented. That is, monitoring rule determination, optimization requirement and policy determination, network/service assurance intent evaluation, and performance deterioration prediction need to be automated.

7.2.3 NR Network UE Throughput Optimization

Based on the framework approach for classification of intelligent network level, the use case of NR network UE throughput optimization defined in clause 6.6.1 achieves intelligent capabilities of level 3:

- Intent management (Human): performed by human
- Awareness (Human & Telecom System): The RAN Manager can automatically collect performance data and generate performance geographic distribution information (for example, grid-based traffic distribution information) based on the collection results.
- Analysis (Human & Telecom System): Based on the manually specified adjustment objectives of locating and optimization solutions, the RAN Manager can automatically detect and identify network performance issues, such as load imbalance, based on network scenarios. Then, the Manager locates and analyzes the detected performance exceptions and potential exceptions to identify the root causes of performance issues.
- Decision (Human & Telecom System): After the RAN Manager generates an optimization solution, the network automatically evaluates the optimization solution, continuously performs iterative optimization, and finally provides the optimal solution.
- Execution (Telecom System): Based on the optimal solution, the RAN Manager delivers configurations and automatically configures the optimal parameter combinations.

Table 7-4 Level evaluation of NR network coverage optimization

Network autonomy level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent management
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary intelligent network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate intelligent network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.						

Level-3 policy driven closed-loop NR network UE throughput optimization automation is achieved in the use case. To achieve evolution to level-4 intelligent networks, the RAN Manager additionally needs to predict potential UE throughput issues and generate UE throughput optimization policies automatically based on the wireless network/service assurance intent. At level 4, intent driven closed-loop coverage optimization automation based on specific service assurance intents for certain scenarios will be implemented.

7.2.4 Root Cause Analysis of Cell Performance Issue

According to the intelligent classification method of network optimization in the above chapter, the intelligent root cause analysis of cell problems is between level 2 and level 3. It realizes the multi-dimensional intelligent association of KPI and MR information on the wireless side of the network, identification of abnormal cell problems, identification of root cause problems, and intelligent classification and perception.

- Intent management (Human): performed by human.
- Awareness (Telecom System): After the automatic input of multi-dimensional feature information, it can then pass through the intelligent analysis engine module, without the need for manual experts to set rules, and can perceive and intelligently output the intelligent classification and identification of abnormal communities and problems.
- Analysis (Human & Telecom System): Combining the KPI and MR information on the wireless side, the intelligent system can initially analyze some causes of problems in the cell and give the root cause analysis results of the cell problems. In the future, network optimization engineer can further analyze and give optimization plan adjustment strategies. It is currently completed in a combination of telecommunications system and manual work.
- Decision (Human): The current decision-making system is mainly based on network optimization experts making final decisions based on the results of perception and analysis. In the follow-up, it will be further combined with the expert system to form an automated decision-making system, which will continue to iterate and optimize, and finally give an optimal decision-making plan. After the RAN domain automatically generates the optimization plan, the network self-evaluates the optimization plan, and iteratively optimizes, and finally gives the optimal plan.
- Execution (Telecom System): The current execution tasks are all automated, and the input data is collected, extracted, analyzed and stored automatically, and automatically input to the artificial intelligence/machine learning processing module for automated operation. The machine learning analysis module can automatically store the analysis results in the database. The optimized execution of the parameters after the decision can be issued and executed through automated procedures.

Table 7-5 Level evaluation of Root cause analysis of cell issue

Network autonomy level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent management
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary intelligent network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate intelligent network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full intelligent network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.						

Subsequent evolution to the 3/4 level smart network, the intelligent root cause analysis system for abnormal communities needs to realize the intelligence and automation of the further analysis part on the current basis, and can be combined with the expert system for full telecommunication system-level analysis and decision-making, and does not rely on manual intervention can automatically and intelligently give decisions, forming an optimized closed-loop automation.

7.2.5 Root Cause Analysis of Alarm

The According to the "mobile communication smart network classification standard", the alarm root cause analysis realizes the following smart network functions.

- Analysis (Human & Telecom System): the network management preprocesses the configuration and alarm data used in algorithm training, and then automatically analyzes the association and causal relationship between alarms according to AI algorithm, identifies the root cause alarm and corresponding derived alarm, and then forms the association and causal relationship rule database. The automatic alarm association rule base needs to be revised again according to the experts' experience. In other words, the alarm root cause analysis process is manual + automatic, and the classification level is between level 2 and level 3.

In order to reach the next level of the analysis process, the root cause analysis process needs to be automated. At present, there is a manual experience rule correction part in the alarm root cause analysis, which can be automatically corrected in the algorithm.

7.3 Summary

The path towards fully intelligent network will be a long-term process of gradual evolution. Right now, most of the use cases in the industry are between level 2 and level 3. In order to evolve to level 4, intent driven closed-loop coverage optimization automation based on specific service assurance intents for certain scenarios needs to be implemented, which including intelligent capability for optimization requirement and policy determination, network/service assurance intent evaluation, and performance deterioration prediction.

In order to achieve this goal, the entire industry needs to have a unified understanding of the intelligent network and its development path. The definition of the intelligent network needs to be continuously clarified through various cases, and most importantly to make the framework being applied in the evaluation of the network. All parties should work together to make a clear definition for each level of intelligent network, and pay attention to the key problems may be found in the practice of use cases, such as insufficient training data, incomplete standardized interfaces, and non-optimized algorithms.

8 Intelligent Network Architecture

8.1 Introduction

In 5G we expect AI/ML to be used at all levels, from an intelligent NE, to intelligent domain management, to cross domain. A multi-level ML capable system needs the ability to use advanced automation to solve specific use cases without jeopardizing the overall functionality. To gradually achieve the goal of a fully intelligent autonomous network and realize AI in mobile networks, hierarchical architecture is a promised solution, in which upper layer is for cross-domain operation and lower layer is for single-domain/network entity (NE) layer operation. ML techniques can be applied to a wide range of use cases in mobile networks. To understand the impact of AI/ML on Mobile Network it is important to identify the requirements for the relevant use cases (for data collection, inference/prediction taken place, actions apply, etc.), from which functional and architectural requirements of a generic AI/ML architecture can be derived.

8.2 Framework Architecture of Intelligent Network

To gradually achieve the goal of a fully intelligent autonomous network and realize AI in Network, without further increasing the network complexity, hierarchical architecture needs to be ensured. A more upper-layer and centralized deployment location indicates a larger data volume, raises higher computing power requirements, and is more suitable to perform cross-domain global policy training and reasoning which does not have real-time requirements.

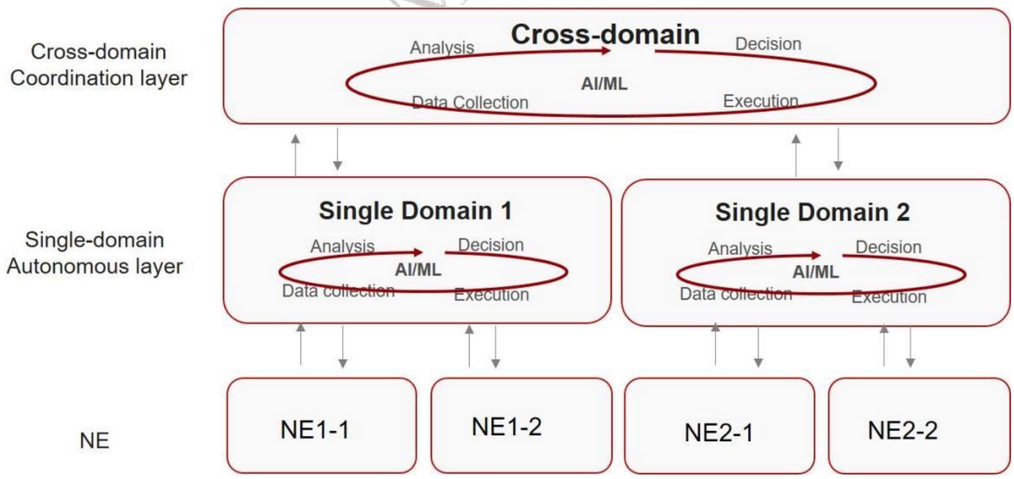

Figure 8-1 General network architecture for intelligent network

For example, cross-domain scheduling and E2E orchestration have high requirements on computing capabilities, require massive cross-domain data, and have low requirements on real-time performance. While at the lower-layer and closer to the end, the domain-specific or entity-specific analysis capability may be possible and real-time performance requirements may be met. The following figure shows a three-layer architecture consisting of the cross-domain orchestration layer, single-domain autonomous layer, and network entity (NE) layer.

8.2.1 Use Cases Classification

ML techniques can be applied to a wide range of use cases in mobile networks. To understand the impact of AI/ML on Mobile Network it is important to identify the requirements for the relevant use cases, from which functional and architectural requirements of a generic AI/ML architecture can be derived. Specifically, the following should be considered:

- The required information (both for ML training and for operational use) and involved network entities.
- The actions to be conducted, and the involved functions and entities in the network.
- The granularity/scale of the action in the network (e.g. cross-domain, domain-specific or entity-specific)
- The time scale of the operation, ranging from near real-time (milliseconds) to minutes to hours and days.

Figure 8-2 illustrates a classification of exemplary use cases along two dimensions - the operational time granularity (typical operating time-scale for a specific solution) and the scope of operation (from network management to network elements). The figure shows a snapshot of the use-cases described in the white paper.

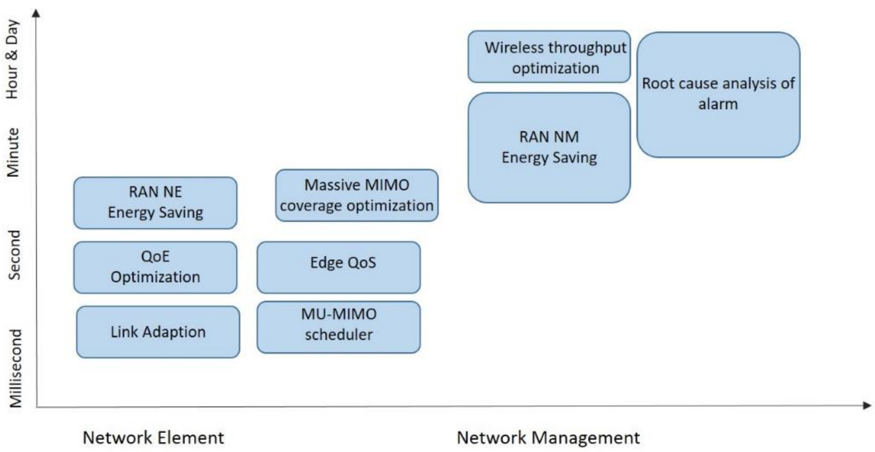

Figure 8-2 classification of exemplary use cases

This categorization enables a mapping of use-case functions to network Element and Network Management, from which (for data collection, inference/prediction taken place, actions apply, and information flow connection the entity) network architecture requirements can be derived.

For example, the use case “QoE Optimization” usually involves a network element function located in gNB and MEC. “QoE Optimization” perform inference/predication and predict the UE’s radio link quality and VR stream data volume in next 10 millisecond. This implies that for this specific case, data collection interfaces and control channel towards RAN and MEC are needed in 10 millisecond level. At the same time, mechanisms to support “QoE Optimization” data analytics and ML inference are needed. MEC and gNB need to carry out the inference policy from “QoE Optimization”.

Different AI/ML use cases have varied time constraints. At the tightest scale, RAN use cases like link adaption MU-MIMO scheduler would have 50 µs–100 µs latency criteria. These are followed by use cases e.g QoE/QoS optimization that need from 10 ms to a few seconds latency criteria. The least demanding in terms of latency are management level use cases, e.g., Massive MIMO coverage optimization, root cause analysis of alarm, that can afford minutes, hours or days of latency. These criteria form an important input to the placement, chaining and monitoring of an ML pipeline.

8.3 Architecture of Typical Use Cases

8.3.1 QoE Optimization

The example of QoE optimization architecture is shown in Figure 8-3. For the “QoE Optimization” use case, the intelligent management function is deployed to manage intelligent applications and provide the data and management channels to application server and BTS. Traffic requirements of the local VR application service are collected and the radio status of UE from the gNB. “QoE Optimization” perform inference/predication. Firstly, “QoE Optimization” predicts the UE’s radio link quality and VR stream data volume in next 10 milliseconds. Then, “QoE Optimization” coordinate VR streaming encoding rate and radio resource scheduled to optimize the user experience and prevent QoE degradation (video stream jitter, mosaic etc.) as the fluctuations of wireless transmission. This implies that data collection interfaces and control channel towards RAN and local service are needed in 10 millisecond level. At the same time, mechanisms to support “QoE Optimization” data analytics and ML inference are needed, e.g. ML model deployment, management, and computer resource allocation etc.   The inference policy will be carried out in MEC and gNB.

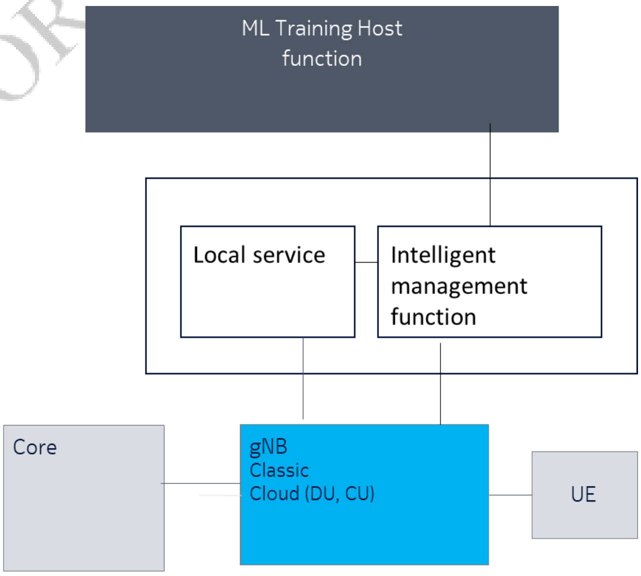

Figure 8-3 QoE optimization architecture (example)

8.3.2 ML-Based MU-MIMO Scheduler

For the ML-based MU-MIMO scheduler use case, Deep Q-Learning or Reinforcement learning is used to find the set of beams that maximize the Q value. Scheduler information e.g. UE's CQI, SINR, MUPF is collected from L2 scheduler per TTI, Usually Inference of Q value will be carried out in DU and complete within 50us-100us to meet the requirement of L2 scheduler. As resource-constrain, DU only host the part of ML function. Q-learning modelling training will be performed in external analytics/ML platform, which collect training data from DU and update model to DU via programmable API for RAN.

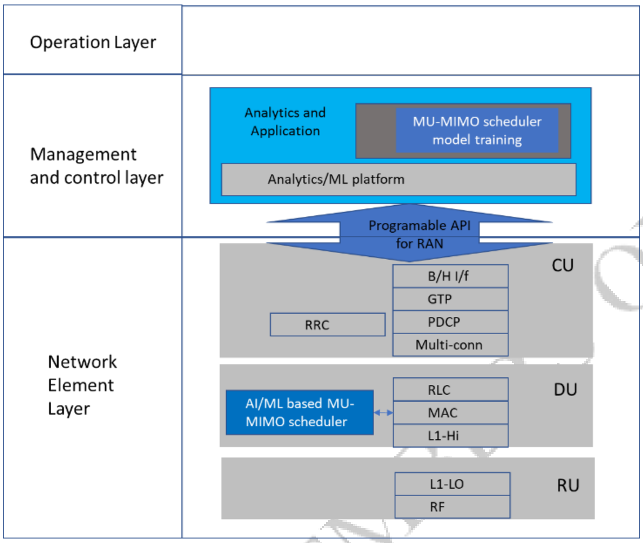

Figure 8-4 ML-based MU-MIMO scheduler architecture

8.3.3 Root Cause Analysis of Alarm

According to the above scheme description, the current alarm root cause analysis scheme mainly focuses on single domain analysis and management at the management and control level. Therefore, several processing modules need to be added in the control layer: data processing, single domain AI platform, model application, as shown in the figure:

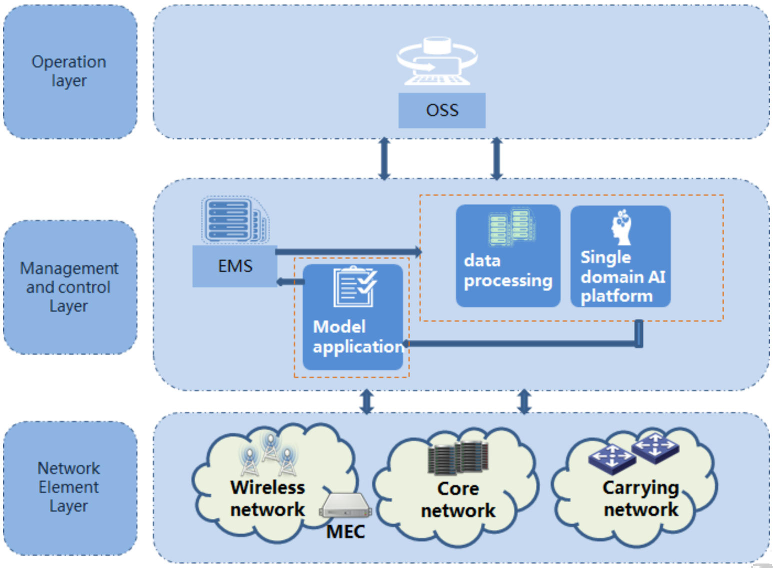

Figure 8-5 Root cause analysis of alarm Architecture

- Data processing: data processing refers to processing the collected original data into training data for algorithm use. Including data storage, data cleaning, data integration, data normalization processing, data security management and other functions.
- Single domain AI platform: single domain AI platform provides AI model training and generation in different scenarios of network business (such as generating association rule model...). It includes feature extraction, model training, algorithm library, model generation, etc.
- Model application: apply the results of model generation to network management system, and evaluate the effect of model application. It includes data acquisition, data preprocessing, model import, model application effect evaluation, etc.

8.3.4 Link Adaptation

This use case is implemented at NE layer, it introduces a machine learning algorithm for steering of the existing link adaptation. The ML algorithm is trained to recognize refined interference scenarios based on the history of the neighbor cell activities and UE signal quality.

The ML models are trained offline. Labelled data is collected from many simulations, lab trials and field trials, the trained ML model are loaded into RAN as part of baseband software.

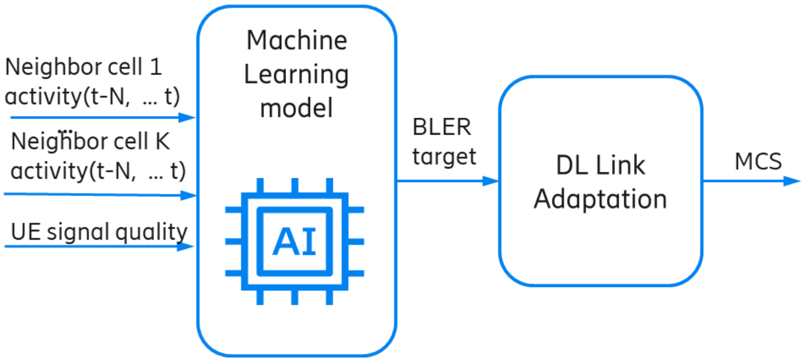

Figure 8-6 Process of AI based DL link adaptation

When the feature is enabled, active cells report statistics over their scheduled PRBs for a short period to a limited number of their neighbour cells, each cell keeps a list of tracked neighbour cells sorted by the signal quality.

The ML algorithm takes actions based on the history of the neighbour cell activity and UE signal quality. The ML algorithm selects BLER target dynamically in time and individually for each UE for a short period of time (sub-seconds). This replaces the constant homogenous BLER target of 10%. So, it periodically adjusts dynamic BLER target for the UE to the current inter-cell interference situation, and DL link adaptation feature selects proper MCS based on the BLER target.

8.3.5 Energy Saving

As shown in Figure 8-7, the cross-domain manager manages the energy-saving service of each domain manager through the service management module. Using the CM, PM, and MR data of each base station, the cross-domain manager generates an energy saving policy based on the AI module and delivers the policy to the base station through the O&M module to implement the policy.

The AI management module in the cross-domain manager manages AI model data in different function modules through the AI data management interface to update and migrate AI model data.

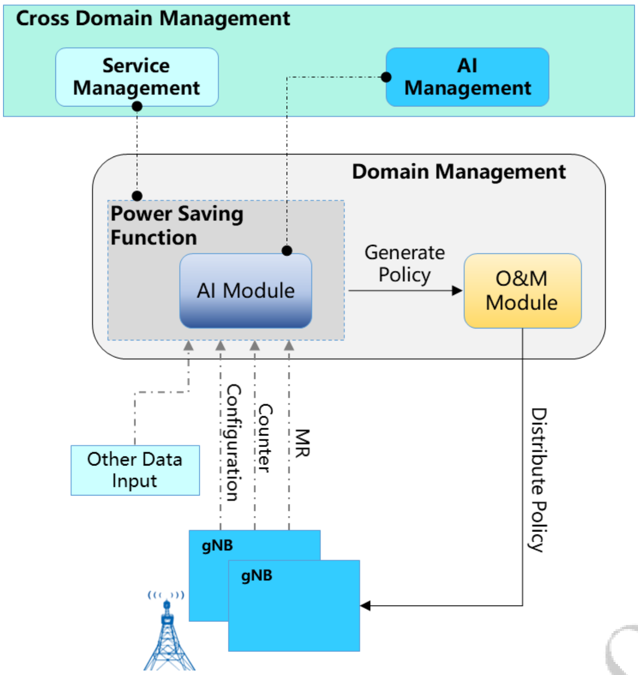

Figure 8-7 Three Layers Energy Saving with AI

8.4 Summary

From the user case architecture presented in 8.3, we can see:

- The deployment of an ML pipeline will span different network layers and domains.
- ML application/user case corresponding requirements (e.g., a traffic classifier requires to be fed with data summaries every x ms) and capabilities of the node (e.g., computing power at the edge) impact distribution of ML entities.
- To support intelligent autonomous network, the network architecture must be extended with data exploitation architectural components. Specifically, it mandates support for flexible and programmable data sources, and deployment of means to exploit real-time, "quasi-real-time" or non-real time information from all kind of sources

9 Intelligent Network Elements

9.1 Introduction

With the rapid development of 5G, the world has entered the rapid development channel of 5G. The low-coverage problem caused by high frequency points greatly increases the density of RAN equipment. According to the general assessment in the industry, the density of 5G equipment is about 2~3 times that of 4G equipment. The operation of wireless access devices on the RAN side directly determines the quality of mobile communication networks. At the same time, the operator will face the O&M network with 2/3/4/5G four generations. The complexity of maintenance and management objects will greatly increase, and the difficulty in fault location and delimitation will be improved. The diversity of 5G ultra-bandwidth, low latency, big connection and other service scenarios impose higher requirements upon the technical capabilities of O&M personnel. With the rapid increase of network scale and the high complexity of service scenarios, the operation mode must be evolved to automatic and intelligent operation.

This chapter describes the functional requirements for network elements from the perspective of intelligence evolution.

9.2 Function Requirements for Network Elements

9.2.1 Multi-Dimensional Data Collection and Reporting

As the main component of a network, NE need to provide richer operation data with more complete dimensions and structures, and provide an automatic collection and reporting mechanism to support the intelligent improvement of cases.

9.2.2 Intelligent Data Modelling

Currently, NE data is more of the traditional design, such as alarms, performance, and configuration. With the increase of data volume and dimensions, the association between data of different dimensions is not high, resulting in the difficulty of intelligent use cases in data cleansing and analysis.

NE needs to model data from the perspective of intelligence to improve the extensive contact of multi-dimensional data.

9.2.3 Data Storage and AI Computing Capability

In terms of data acquisition convenience, network elements have natural advantages, for example, richer user-level and service-level data. In addition, underlying data such as the physical layer can be used to improve the accuracy and timeliness of intelligent use cases. With the enhancement of storage and computing capabilities, it is also possible to store and calculate AI data on the NE side.

NE need to provide small-scale data storage capabilities and AI computing capabilities to support the implementation and deployment of intelligent use cases with real-time requirements.

9.2.4 Intelligent Feedback and Closed-loop Control

After intelligent use cases of NE are deployed, a feedback and closed-loop control mechanism needs to be established with the EMS, for example, optimization parameter delivery and real-time calculation feedback, to implement intelligent use case management and effect optimization.

9.2.5 AI Model Interactionn

NE needs to support AI model interaction with intelligent management function, including interactive interfaces for downloading, loading, rolling back, and querying AI models.

10 Intelligent Network Management

10.1 Introduction

In a typical operator application scenario, the domain manager interconnects with the NMS through an open API. The NMS delivers the network coverage optimization objectives and areas to be optimized to the domain manager. The domain manager sends the final optimization result and optimization advice of each round to the NMS of the operator.

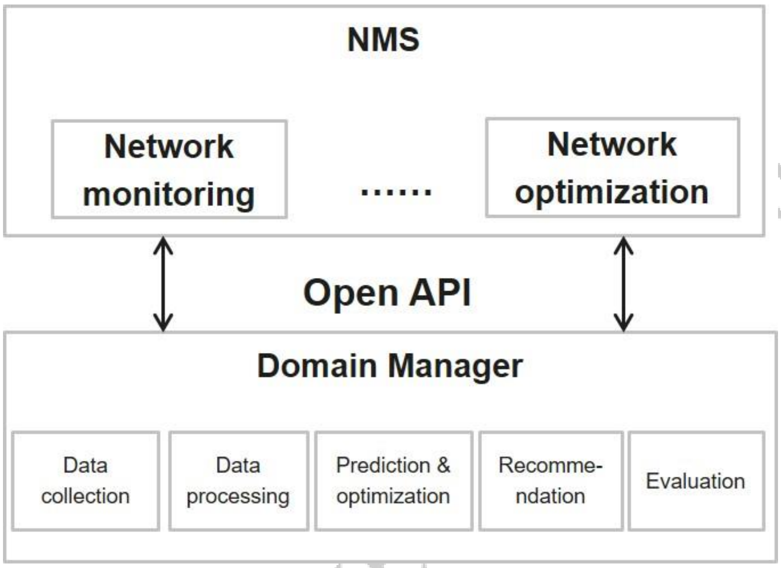

Figure 10-1 Network framework in a typical operator application scenario

10.2 Function Requirements for Network Management

In order to achieve intelligent network, domain manager shall have the function network management and control, and have the following functions:

10.2.1 Data Collection and Reporting

The domain manager shall automatically collect performance data, for example coverage, traffic data. And it shall be the local knowledge base and AI inference framework for single domain network.

10.2.2 Data Analysis and Modelling

Based on the manually specified adjustment objectives of locating and optimization solutions, the domain manager shall automatically detect and identify network performance issues, such as weak coverage, load imbalance, based on network scenarios. Furthermore, the domain manager shall identify the root causes of performance issues.

The domain manager shall have the function of data processing and model training. It shall generate optimization solution for single domain based on different scenario and deliver to the network element.

10.2.3 Network Management and Control

After an optimization solution is generated, the domain manager shall automatically evaluate the solution, continuously perform iterative optimization, and finally provide the optimal solution.

For example, based on the optimal solution, the domain manager shall deliver configurations and automatically configure the optimal parameter combinations.

11 Conclusion and Recommendation

- Intelligent network will be an important enabler for 5G network evolution and enhancement.
- SDOs and industry parties are recommended to work together on network intelligence to help the industry to progress in a coordinated way.
- The use cases of network intelligence can be categorized based on the dimension of full life cycle and main functional entities.
- Full intelligent network will be a long-term target. Intelligent network level describes the level of intelligence capabilities of the network and it is beneficial for the industry on roadmap planning. Currently, most of the use cases in the industry are between level 2 and level 3.
- The deployment of an ML pipeline will span different network layers and domains. The network architecture must be extended with data exploitation architectural components.
- Network elements need to enhance the capabilities and interfaces to meet the function requirements including multi-dimensional data collection and reporting, intelligent data modelling, data storage and AI computing, intelligent feedback and closed-loop control, etc.
- Network management need to enhance the capabilities and interfaces to meet the function requirements including data collection and reporting, data analysis and processing, ML model training, distributing and update, network management and control, etc.