GTI 5G Autonomous Network White Paper

http://gtigroup.org/

GTI 5G Autonomous Network White Paper

Version:	V2.0
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: Autonomous Network
Task	Autonomous Network Level (ANL) Autonomous Network Architecture (ANA) Autonomous Network Elements (ANE) Autonomous Network Management (ANM)
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 White paper. 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.
Dec. 21,2021	32nd GTI Online Workshop	V2.0	From this 2nd version, the title of this white paper evolves into “GTI 5G Autonomous Network White Paper” according to the corresponding project renaming. Based on the first version, corresponding contents are updated, relevant information and data of use cases are up to date, the “essential capacities” are described and analyzed in details, in the meantime, some topics of future study are also introduced for the industry further discussion.

Table of Contents

1 Executive Summary

This is the second version of GTI 5G Intelligent Network White Paper, whose first version has been published in November 2020. And as industry & technology developing and evolving, the title of this version evolves into "**GTI 5G Autonomous Network White Paper**".

Autonomous Network is defined as “telecommunication system (including management system and network) with autonomy capabilities which is able to be governed by itself with minimal to no human intervention” in 3GPP TS28.100 and defined as “a system of networks and software platforms that are capable of sensing its environment and adapting its behavior accordingly with little or no human input” in TM Forum IG1230. According to these definitions, automatic and intelligent mechanisms are the essential enablers of autonomous networks to minimize human intervention. With the development of network systems and evolution of AI technology applications, self-X properties (the abilities to monitor, operate, recover, heal, protect, optimize, and reconfigure themselves) are expected to be achieved in autonomous network.

In this white paper, corresponding contents in the previous version have been updated, relevant information and data of use cases are up to date, in the meantime, some topics of future study are also introduced for the industry further discussion.

It covers current standardization and industry status, updated use cases, essential capacities, autonomous network level, autonomous network architecture, autonomous network elements and autonomous network management and further study for autonomous network.

The information of standardization status is described in general, and following, the updated use cases have been introduced for further study. The categories of use cases have been updated based on the dimension of full life cycle and main functional entities.

All the capacities of autonomous network have been refined into four essential capacities, i.e. autonomous perception, autonomous prediction, autonomous diagnosis and autonomous control. Each of these essential capacities has been illustrated and described as a clause, in which the relevant use cases have been also illustrated to make a clearly description and understanding for each essential capacity.

In this version of white paper, autonomous network level has been updated depending on the current R&D achievements, the progress of standardization and the evaluation of commercial applications.

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

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

In addition, further studies for autonomous network have been explored in the topics of "trust in autonomous networks" and "life cycle management" are discussed so that to attract the attention and further study of the industry. And this topic will be opened for all the industry partners to contribute to make some further study and discussion of relevant cutting-edge topics.

2 Reference

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

[1] ITU-T Y.3170, “Requirements for machine learning-based quality of service assurance for the IMT-2020 network”, 2018.
[2] ITU-T Y.3172, "Architectural framework for machine learning in future networks including IMT-2020", 2019.
[3] ITU-T Y.3173, "Framework for evaluating intelligence levels of future networks including IMT-2020", 2020.
[4] ITU-T Y.3174, "Framework for data handling to enable machine learning in future networks including IMT-2020", 2020.
[5] ITU-T Y.3175, “Functional architecture of machine learning-based quality of service assurance for the IMT-2020 network”, 2020.
[6] ITU-T Y.3157, “Architectural framework for artificial intelligence-based network automation for resource and fault management in future networks including IMT-2020”, 2021.
[7] ITU-T Deliverable “Trust in Autonomous Networks”, https://www.itu.int/en/ITU-T/focusgroups/an/Pages/default.aspx
[8] ITU-T Deliverable “High level architecture framework for Autonomous Networks”, https://www.itu.int/en/ITU-T/focusgroups/an/Pages/default.aspx
[9] ITU-T Supplement Y.Supp-AN-Use Cases, “Use cases for Autonomous Networks”.
[10] ITU-T Y.DTN-ReqArch, "Requirements and Architecture of Digital Twin Network".
[11] ITU-T Y.IBNMO "Intent-based network management and orchestration for network slicing in IMT-2020 networks and beyond".
[12] ITU-T M.3080, “Framework of artificial intelligence enhanced telecom operation and management (AITOM)”, 2021.
[13] ITU-T M.IL-AITOM, “Intelligence Levels of AI enhanced Telecom Operation and Management”.
[14] ITU-T M.RLA-AI, “Requirements for Log Analysis with AI-enhanced Management System”.
[15] ITU-T M.RMNOC-AI, “Requirements for the management of network operation cost within AITOM in telecom operational aspects”.
[16] ITU-T M.RWOP-AI, “Requirements for work orders processing in Telecom Management with AI”.
[17] 3GPP TR 28.810, "Study on concept, requirements and solutions for levels of autonomous network".
[18] 3GPP TS 28.100, "Management and orchestration; Levels of autonomous network".
[19] 3GPP TS 28.541, "Management and orchestration; 5G Network Resource Model (NRM); Stage 1".
[20] 3GPP TS 28.552, "Management and orchestration; 5G performance measurements".
[21] 3GPP TS 32.422, "Telecommunication management; Subscriber and equipment trace".
[22] 3GPP TS 28.307, "Telecommunication management; Quality of Experience (QoE)."

[23] 3GPP TR 28.812, "Telecommunication management; Study on scenarios for Intent driven management services for mobile networks."
[24] TS 28.535, "Management and orchestration; Management services for communication service assurance; Requirements."
[25] 3GPP TR 28.809, "Study on enhancement of management data analytics."
[26] TS 28.313, "Management and orchestration; Self-Organizing Networks (SON) for 5G networks."
[27] 3GPP TS 28.555, "Management and orchestration; Network policy management for 5G mobile networks; Stage 1."
[28] 3GPP TS28.557, "Management and orchestration; Management of Non-Public Networks (NPN); Stage 1 and stage 2."
[29] 3GPP TS 28.541, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3."
[30] 3GPP TR 37.816, "Study on RAN-centric data collection and utilization for LTE and NR".
[31] 3GPP TS 38.314, "NR; Layer 2 measurements".
[32] 3GPP TS 38.300, "NR; Overall description; Stage-2".
[33] 3GPP TS 37.320, "Minimization of Drive Tests (MDT); Overall description; Stage 2".
[34] 3GPP TS 38.306, "NR; User Equipment (UE) radio access capabilities".
[35] 3GPP TS 38.331, "NR; Radio Resource Control (RRC); Protocol specification".
[36] 3GPP TR 23.791, "Study of enablers for Network Automation for 5G".
[37] 3GPP TS 23.288, "Architecture enhancements for 5G System to support network data analytics services".
[38] 3GPP TR 23.700-91, "Study on Enablers for Network Automation for 5G-phase2".
[39] 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
[40] TMF, "Autonomous Networks: Empowering digital transformation for smart societies and industries", https://www.tmforum.org/resources/whitepapers/autonomous-networks-empowering-digital-transformation-for-smart-societies-and-industries/
[41] TMF, "Cross-Industry Autonomous Networks - Vision and Roadmap", https://www.tmforum.org/resources/how-to-guide/ig1193-cross-industry-autonomous-networks-vision-and-roadmap-v1-0/
[42] TMF, "Business requirements & framework", https://www.tmforum.org/resources/standard/ig1218-autonomous-networks-business-requirements-and-architecture-v2-0-0/
[43] TMF, "Autonomous Networks Technical Architecture", IG1230 Autonomous Networks Technical Architecture v1.1.0 | TM Forum
[44] TMF, "Intent in Autonomous Networks", https://www.tmforum.org/resources/how-to-guide/ig1253-intent-in-autonomous-networks-v1-0-0/
[45] ETSI GS ZSM 002, "Zero-touch network and Service Management (ZSM); Reference Architecture". 2019.08.
[46] ETSI GS ENI 001, "Zero-touch network and Service Management (ZSM); Requirements based on documented scenarios", 2019.10.
[47] ETSI GS ENI 005, "Zero-touch network and Service Management (ZSM); Means of

Automation", 2020-05.
[48] ETSI GS ENI 007, "Zero-touch network and Service Management (ZSM); Terminology for concepts in ZSM," 2019-08.
[49] ETSI TS 103.195-2 (2018-05): "Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management ", https://www.etsi.org/deliver/etsi_ts/103100_103199/10319502/01.01.01_60/ts_10319502v010101p.pdf
[50] ETSI TS 103 194 (2014-10): "Network Technologies (NTECH); Autonomic network engineering for the self-managing Future Internet (AFI); Scenarios, Use Cases and Requirements for Autonomic/Self-Managing Future Internet", https://www.etsi.org/deliver/etsi_ts/103100_103199/103194/01.01.01_60/ts_103194v010101p.pdf
[51] TS 103.195-2, "Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management," 2018-05.
[52] TS 103 194, "Scenarios, Autonomic Use Cases and Requirements for Autonomic/ Self-Managing Future Internet".
[53] CCSA TC1, "Technical report of telecommunication network planning application based on artificial intelligence".
[54] CCSA TC5, "Study on grading method for intelligent capability of mobile networks".
[55] CCSA TC5, "Intelligent 5G Radio".
[56] CCSA TC7, "Technical specification for intelligent level of mobile network management and operation".
[57] CCSA TC7, "Functional architecture of intelligent information & communication operation management system".
[58] ETSI GS ENI 005, "Experiential Networked Intelligence (ENI); System Architecture, V1.1.1", 2019-09.
[59] Qing Zhang, “AI based intelligent recognition of 5g base station energy saving scenarios”, Aug 2019.
[60] Jianfeng Lei, “Research on network traffic prediction based on Neural Network”, May 2008.
[61] Zhirong Zhang, “Research on energy saving technology of 5G base station based on AI”, Oct 2019.
[62] Pang, Eugene, Ericsson, "How AI-powered capacity planning will impact network expansion decisions", Dec 2020, https://www.ericsson.com/en/blog/2020/12/ai-and-network-capacity-planning-decisions
[63] Ericsson, "Generating actionable insights from customer experience perception", Sept 2016, https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/generating-actionable-insights-from-customer-experience-perception
[64] Ericsson, "white paper of accelerating-the-adoption-of-ai-in-programmable- 5g-networks," https://www.ericsson.com/en/reports-and-papers/white-

papers/accelerating-the-adoption-of-ai-in-programmable-5g-networks

[65] ITU-T Y.3052, Recommendation ITU-T Y.3052 (2017), “Overview of trust provisioning in information and communication technology infrastructures and services”.

3 Abbreviations

Abbreviation	Explanation
3D	Three-Dimensional
3GPP	3rd Generation Partnership Project
AAU	Active Antenna Unit
AI	Artificial Intelligence
ANL	Autonomous Network Level
ANR	Automatic Neighbor Relation
AON	Autonomous Network
API	Application Programming Interface
BTS	Base Transceiver Station
CCO	Coverage and Capacity Optimization
CCO	Coverage and Capacity Optimization
CCU	Cell Center User
CE	Cell Edge
CEU	Cell Edge User
CM	Configuration Management
CQI	Channel Quality Indication
CSP	Communication Service Providers
DL	Down Link
DT	Drive Test
E2E	End to End
EMS	Network Element Management System
eNB	E-UTRAN NodeB
EOMS	Electric Operation Maintenance System
E-RAB	Evolved Radio Access Bearer
FG-AN	Focus Group on Autonomous Networks
FP	Frequent Pattern
FP-G	Frequent pattern Growth
GANA	Generic Autonomic Networking Architecture
GSMA	Global System for Mobile Communications Association
GTI	Global TD-LTE Initiative
HAS	Http Adaptive Streaming
HO	Hand Over
KNN	K-Nearest Neighbor
KPI	Key Performance Indicator
LB	Load Balance
LCM	Life Cycle Management
LTE	Long Term Evolution
MANO	Management and Orchestration
MDT	Minimization of Drive Tests

Abbreviation	Explanation
MEC	Multi-Access Edge Computing
MIMO	Multiple-input multiple-output
ML	Machine Learning
MR	Measurement Report
MU-MIMO	Multi-User Multiple-Input -Multiple-Output
NE	Network Element
NFV	Network Function Virtualization
NMS	Network Management System
NR	New Radio
NWDAF	Network Data Analytics Function
O&M	Operation and Maintenance
OODA	Observe, Orient, Decide, Act
OPEX	Operating Expense
PDU	Protocol Data Unit
PM	Performance Management
PnP	Plug and Play
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
SDO	Standards Development Organizations
SG	Study Group
SON	Self-Organizing Network
TCO	Total Cost Of Ownership
TTI	Transmission Time Interval
UE	User Equipment
UPF	User Plane Function
VNF	Virtual Network Function
VR	Virtual Reality

4 Introduction

This white paper is the second version of GTI 5G Intelligent Network White Paper, and from this version, the title of this white paper evolves into **"GTI 5G Autonomous Network White Paper"**. Just as the name expresses that, this white paper mainly focuses on 5G autonomous network, including but not limited to the status of global standardization and industry, use cases analysis, essential capacities, autonomous network levels, autonomous network architectures of the representative use cases, and general function requirements on autonomous network elements and autonomous network management, in the meantime, further studies of Autonomous Network. Further study and descriptions are detailed in the version of white paper based the original one.

Sincere thanks to all the contributors and the supporters for their hard work in writing this white paper. 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, CICT, Ericsson, Huawei, Nokia, ZTE
- **Chapter 6 Use Cases**
  China Mobile, CICT, Ericsson, Huawei, Nokia, ZTE
- **Chapter 7 Essential Capacities**
  CICT, Ericsson, Huawei, Nokia, ZTE, China Mobile
- **Chapter 8 Autonomous Network Level**
  Huawei, CICT, Ericsson, China Mobile
- **Chapter 9 Autonomous Network Architecture**
  Nokia, China Mobile
- **Chapter 10 Autonomous Network Elements**
  ZTE, China Mobile
- **Chapter 11 Autonomous Network Management**
  Huawei, CICT, China Mobile
- **Chapter 12 Further Study**
  China Mobile, Ericsson

Special thanks to the following contributors for writing the white paper.

- **China Mobile**
  Ms. Xiaojia SONG, Dr. Xi CAO, Mr. Zhuowen GUAN, Mr. Guangyu LI, Mr. Xiaolei HUA, Ms. Xiaobin GUAN, Mr. Huan YANG, Mr. Xin CHEN, Dr. Lingli DENG, Ms. Yanping LIANG, Dr. Lin ZHU, Mr. Li YU, Mr. Xiangyang YUAN, Dr. Junlan FENG

- **CICT**
  Ms. Xuan ZHANG
- **Ericsson**
  Ms. Xiaomei LIU, Mr. Bo GUAN, Mr. Liyu LIU, Mr. Christoffer STUART, Mr. Meng LAI
- **Huawei**
  Ms. Dannie CHEN, Mr. Ruiyue XU, Mr. Zhi Jie XU, Mr. Chunhong YUAN
- **Nokia**
  Mr. Zhuyan ZHAO, Ms. Fangfang GU
- **ZTE**
  Mr. Pan LI, Mr. Honghui KANG, Ms. Hong FENG, Ms. Yin GAO, Mr. Zijiang ZHANG, Mr. Yi DING

This is the second version of the white paper, it will be continuously updated according to the research and development progress.

5 Standardization and Industry Status

5.1 Motivation and Overview

The Autonomous Networks are on their way to broader application, further studying and higher autonomous level. All the communication standards development organizations (SDO) and industry parties are actively producing or participating in relevant standards, recommendations, technical reports, white papers, survey and so on. As following, the general status of SDOs and industry parties has been shown.

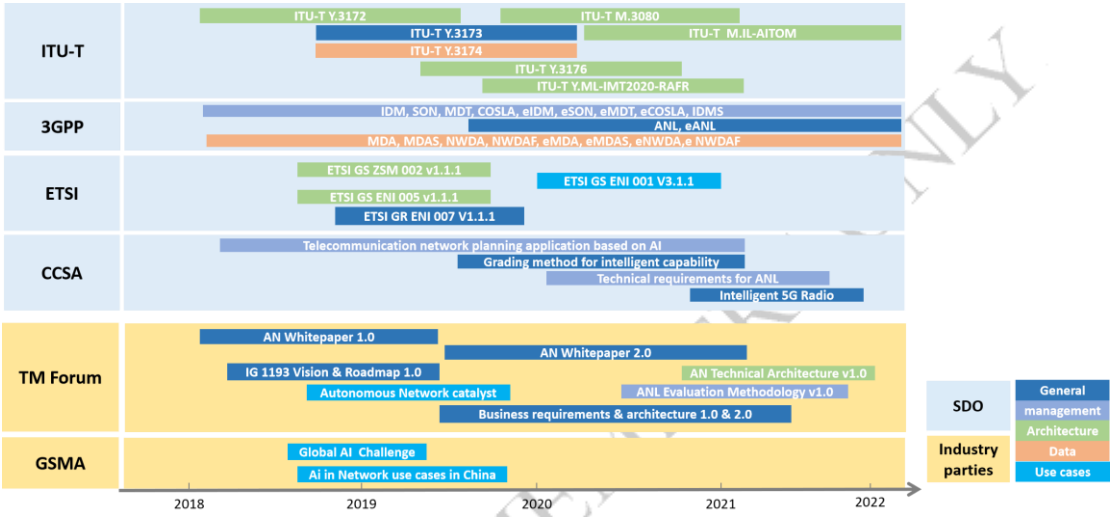


Figure 5-1 General status of SDOs and industry parties

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

5.2 Activities in ITU-T

In ITU-T, Focus Group on Autonomous Networks (FG-AN), Study Group 13 (SG13: future networks, with focus on IMT-2020, cloud computing and trusted network infrastructures) and Study Group 2 (SG2: operational aspects), are the groups which are studying autonomous networks relevant topics.

- Focus Group on Autonomous Networks (FG-AN) has been established in April 2021 and its parent group is SG13. The primary objective of the Focus Group is to provide an open platform to perform pre-standards activities related to this topic and leverage the technologies of others where appropriate. Deliverables including "trust in AN" [7], "High-level architecture of AN" [8], "Use cases of AN" [9] and PoC of AN" are studying and the deliverable of "Use cases of AN" is already delivered to ITU-T SG13, in the meantime, other deliverables are still under study.

- ITU-T SG13: SG13 is the study group of ITU-T which focus on "future networks, with focus on

IMT-2020, cloud computing and trusted network infrastructures", and the topic related to autonomous networks came into study since 2017. Recommendations ITU-T Y.3170 [1], ITU-T Y.3172 [2], ITU-T Y.3173 [3], ITU-T Y.3174 [4], ITU-T Y.3175 [5], ITU-T Y.3177 [6], etc. have already published. The work items Y.DTN-ReqArch [10] and Y.IBNMO [10] are still in studying.

- ITU-T SG2: SG2 mainly focuses on "operational aspects", and in which Recommendation M.3080 [12] has been published. And work items M.IL-AITOM [13], M.RLA-AI [14], M.RMNOC-AI [15] and M.RWOP-AI [16] are on the way.

5.3 Activities in 3GPP

Self-Organizing Network (SON) is introduced by 3GPP since 4G era. Automation of some network planning, configuration and optimization processes via the use of SON functions (e.g. ANR, PnP, CCO) can help the network operator to reduce operating expense (OPEX) by reducing manual involvement in such tasks. In 5G era, Autonomous Network (AON) is formally introduced and defined in 3GPP, which describes the telecommunication system (including management system and network) with autonomy capabilities which is able to be governed by itself with minimal to no human intervention. Following are some standardization for autonomous networks for 5G defined in 3GPP:

- Autonomous Network Level (ANL) describes the level of autonomy capabilities in the autonomous network to improve the efficiency for network management and control. 3GPP SA WG5 started the study item "Study on autonomous network levels" [17] from August 2019 and published in September 2020, which output the concept, dimension, framework and typical use cases for classification of autonomous network level. Since June 2020, a new work item "Autonomous network levels" [18] was approved and started, which provides the concept and framework for evaluating autonomous network levels, and delivers the generic workflow, classification of autonomous network levels and corresponding requirements for the scenarios related to network and service planning, deployment, maintenance and optimization.
- Since Release 16, 3GPP SA WG5 has started a series of standard features to continuously promote the autonomous network and enable different levels of autonomous networks. Including:
    - A series of standard projects for the basic 5G management features (mainly focus on the perception and Execution management tasks) are started since Rel-15 to enable the autonomous network level 1 and autonomous network level 2. For example, 5G NRM features [19], 5G performance measurements and KPIs [20], 5G MDT/trace management [21], QoE Measurement Collection [22], etc.
    - A series of standard projects for the 5G automation or intelligence management features (mainly focus on the Analysis, Decision and Intent handling) are introduced since Rel-16 to enable the autonomous network level 3 and autonomous network level 4. For example, "Intent driven management service for mobile network" [23], "Closed loop SLS Assurance" [24], "Management Data Analytics Service (MDAS)"

[25], "Self-Organizing Networks (SON) for 5G networks"[26], "Network policy management for 5G mobile networks" [27], etc.

- A series of standard projects for the management features to support new vertical service are introduced since Rel-16 to enable the autonomous network applicable for 5GtoB. For example, "Management of non-public networks" [28], "Management Aspects of 5G Service-Level Agreement" [29], etc.

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 [30] and "SON/MDT support for NR" standard project [31]-[35], 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 [36]-[38]. 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 TMF

TMF was one of the first organizations to recognize the powerful value of autonomous networks in the digital transformation business. From 2019, TMF launched the "Autonomous Network Project" and successively released a series of autonomous network white papers, covering TMF's research content on the key aspects of autonomous network:

- AN white paper 1.0 (2019) [39], AN white paper 2.0 (2020) [40], outlining the motivation, vision, framework and use cases for autonomous network.
- IG1193 AN Vision & Roadmap v1.0.1 (2019) [41], serving as the general guideline for pertinent work streams and work items.
- IG1218 AN Business Requirements and Framework v2.0.0 (2021) [42], including business requirements, architecture, capabilities, and related key metrics for measuring autonomous levels as well as life cycle of AN services.
- IG1230 AN Technical Architecture v1.1.0 (2021) [43], as a technical guide for implementation of AN.
- IG1253 Intent in Autonomous Networks v1.0.0 [44], a guide on how to introduce intent handling in the networks to increase autonomy and enable intelligent decision making by the network that aligns with business interests.

TMF's ambition is to create an industry standard framework for Autonomous Networks which is use case driven, based on autonomous domains and intent driven interactions.

For some of use cases, TMF also established the catalyst project Catalysts, which proved the feasibility of these use cases in stages. At the same time, TMF has also developed open APIs for autonomous networks, and is cooperating with other standards organizations such as ETSI to discuss how to further implement API development and use in accordance with ETSI-related

standards and specifications.

5.5 Activities in ETSI

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

- ISG ZSM (ISG Zero-touch Network and Service Management), was formed with the goal to introduce a new end-to-end architecture and related solutions that will enable automation at scale and at the required minimal total cost of ownership (TCO), as well as to foster a larger utilization of artificial intelligence (AI) technologies. The ZSM end-to-end architecture framework [45] has been designed for closed-loop automation and optimized for data-driven machine learning and AI algorithms. The architecture is modular, flexible, saleable, extensible and service-based. It supports open interfaces as well as model-driven service and resource abstraction. Closed loops (e.g. using the OODA model of Observe, Orient, Decide, Act) enable automation of management tasks and allow e.g. continuous self-optimization, improvement of network and resource utilization, and automated service assurance and fulfillment. The ISG ZSM is currently working to specify the closed-loop automation enablers, including automatic deployment and configuration of closed loops, means for interaction between closed loops (for coordination, delegation and escalation), use of policies, rules, intents and/or other forms of inputs to steer their behavior, etc. In addition, the ISG is working to specify closed-loop solutions for particular end-to-end service and network automation use cases, based on the generic enablers and ZSM architectural elements for closed loops as defined in ETSI GS ZSM 002 [45].

- Within ISG NFV (Network Function Virtualization), AI is being considered as a tool that eventually becomes part of the Management and Orchestration (MANO) stack. NFV virtualization is not explicitly considering AI, except in requirements to properly feed data and collect actions from AI modules: Although NFV-MANO has already been equipped with fundamental automation mechanisms (e.g., policy management), it is still necessary to study feasible improvements on the existing NFV-MANO functionality with respect to automation [...] to investigate the feasibility on whether those automation mechanisms can be adapted to NFV-MANO during the NFV evolution to cloud-native.

- The ETSI group ISG ENI designs “Experiential Networked Intelligence” based on data collection and processing using closed loop decision-making. The specification ETSI GS ENI 001 [46] demonstrates a number of use cases on service assurance, fault management and self-healing, resource configuration, performance configuration, energy optimization, security and mobility management. The specification ETSI GS ENI 005 [47] shows as a functional architecture how the data is collected, normalized and recursively processed to extract knowledge and wisdom from it. This data is used for decision-making and the results are returned to the network, where the behavior is continually monitored. The requirements document ETSI GR ENI 007 [48] on network classification of AI details the use of AI in a network into six stages, from "No AI" to "full AI" deployment. Clearly, no network is at either extreme of the six stages. ISG ENI is specifying training methods in document ETSI GS ENI 005

version 2 [47]. Training is often made with big data. Learning is the method used by the AI system to extract knowledge from the training data. Learning can take many forms: dictionary learning, rule-based learning, federated learning, supervised learning, reinforcement learning and "pure machine" unsupervised learning, or combinations of these. Training can be centralized, federated, umbrella-like or distributed peer-to-peer. An AI system that is trained and has learning in a particular field (e.g. image recognition, eHealth, networking and resource management, IoT, robotics, etc.) may continually adapt with further online learning, or may have offline learning to refresh its perception of the situation (re-training).

- TC INT Core Network and Interoperability Testing group created TS 103.195-2 for the Generic Autonomic Network Architecture (GANA) [49] and TS 103 194 "Scenarios, Autonomic Use Cases and Requirements for Autonomic/Self-Managing Future Internet"[50]. The optimization can be categorized as:
    - Actions that are performed on network configuration parameters or network resources, e.g. transmission power, antenna tilt, routing policies, bandwidth allocation.
    - Actions that are performed on the network structure, e.g. adding/removing network elements (either physical or virtualized instances). These actions imply configuration changes in order to accommodate the structural change.
- TC INT specifications consider events that can trigger a network to dynamically change network properties. Events vary depending on the specific AI systems deployed in the network and the level where they operate, external or internal to the network. These events can occur in a chain-like fashion, e.g. policy change can trigger several secondary events in lower level functional units.

ETSI ISG SAI is the first technology standardization group focusing on securing AI. The intent of the ISG SAI is to address 3 aspects of AI in the standards domain: Securing AI from attack (e.g. where AI is a component in the system that needs defending), mitigating against AI (e.g. where AI is the 'problem' or used to improve and enhance other more conventional attack vectors), and using AI to enhance security measures against attack from other things (e.g. AI is part of the 'solution' or used to improve and enhance more conventional countermeasures). There are 5 Active Work Items in SAI: Securing AI Problem Statement, AI Threat Ontology, Data Supply Chain, Mitigation Strategy Report and Security Testing of AI.

5.6 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 [53] on telecommunication network application based on artificial intelligence from July 2018. The research is focusing on network planning and application based on AI. The output report will mainly cover network planning needs, network architecture, data training, and network evolution.

- CCSA TC5 started the research project [54] on intelligence levels of mobile networks in November 2018 and finished in August 2019, and [55] on 5G Radio technology and 5G base station from November 2020 and finished in December 2021. The output of the research covers the methods for evaluating intelligence levels for 5G mobile networks, and typical use cases as well as fundamental functions for classification of intelligence level and potential influence between the 5G mobile network architecture and intelligence levels.
- CCSA TC7 started the standard specification [56] in January 2020, and 错误!未找到引用源。 [57] from July 2020 to December 2022. The work target is to define the classification concept, general evaluation method and fundamental principle for AI technology used in 5G network management and operation.

5.7 Activities in GSMA

Industry bodies such as the GSMA, TM Forum and others are taking active action to explore and promote collaboration on the topic of autonomous networks among SDOs, operators, vendors and other industry players.

At the GSMA, AI and autonomy is one of the themes of the "Network of the Future. "2020 On February 25, 2020 the Global System for Mobile Communications Association (GSMA) announced the winners of the 2020 Global Mobile Awards (GLOMO Awards), with 35 awards in 9 categories.2020 In March, the GSMA, in conjunction with the China Academy of Information and Communications Technology, released the "5G Vertical Industry Application Cases in China In March 2020, the GSMA and the China Academy of Information and Communications Technology (CAICT) released the "China 5G Vertical Industry Application Cases" report. From June 30-July 2, 2020, the GSMA launched the first GSMA Thrive-everything is bright online exhibition on 5G and networking. The GSMA Thrive - The Future of Internet of Things and Security will be discussed at the show. The GSMA Artificial Intelligence Enabled Security Task Force was established at this forum. On October 12, 2020, the "5G Millimeter Wave Industry Summit" hosted by GSMA with the theme of "Unleashing the Full Potential of 5G" was officially launched at the MWC21 Shanghai. The White Paper on 5G Millimeter Wave Technology was officially released at the GSMA-sponsored "5G Millimeter Wave Industry Summit". The White Paper points out that 5G millimeter wave has abundant frequency resources and is an inevitable direction for the evolution of mobile communications technology as the demand for bandwidth continues to increase and the communications spectrum continues to extend to higher spectrum. The theme of the show was "Harmony and Coexistence", and it was conducted through a combination of online platform and offline activities, presenting innovative products and technological breakthroughs in various fields such as 5G practical applications, AI, AIoT, IoT, smart home, smart phones, etc. for the whole communication industry.

5.8 Summary

As the development of technologies and the evolution of networks, autonomous networks 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,

general process, architecture, data and application use cases. There are still a serious of consensus of autonomous networks needed to be achieve through SDO and industry parties.

6 Use Cases

6.1 Introduction

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

This white paper categorizes the autonomous 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, network and service operation, and the dimension of main functional entities i.e. autonomous network element and autonomous 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.
- **Network and Service Operation:** process of network and service configuration depending on customer's requirements, service assurance and monitoring so that to timely discover related problem(s) and quickly respond.

6.2.2 Functional Entity Dimension

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

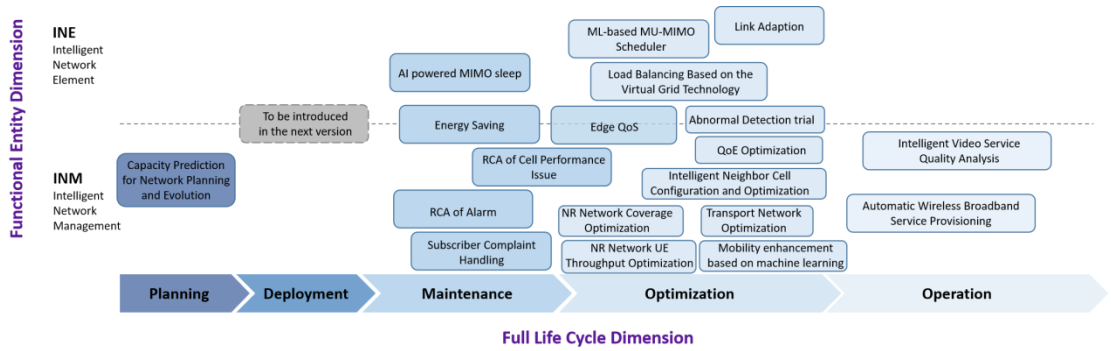


Figure 6-1 Classification of use cases

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. In the meantime, use cases of planning and operation have been updated, use cases of network service deployment would be introduced in the next version.

6.3 Use Case of Network and Service Planning

6.3.1 Capacity Prediction for Network Planning and Evolution

6.3.1.1 Background

The prediction capability is the ability to forecast time series, long- and short-term, or to predict a best action, or best input to an action, based on current environment. These abilities are necessary at all levels of the network. To Capacity prediction, long-term capacity planning can utilize long time-series prediction to maximize use of already installed base and the return on future investment in network capabilities.

Capacity Planning is a solution that accurately predicts the radio resources consumption. The solution is based on automated machine learning models and allows a precise identification of capacity problems and optimal preventative actions in what-if scenarios.

6.3.1.2 Solution Overview

Capacity Planner is a solution that gives CSP the unique ability to better understand the CAPEX and OPEX impact of introducing new services, and the general network dimensioning process.

To achieve this goal, Capacity Planner is based on a combination of Traffic Forecasting with Utilization KPI prediction based on state-of-the-art Machine Learning techniques. This enables the detection of capacity problems in a proactive way, leading to optimum bottleneck identification in a pre-defined time frame as opposed to the traditional way of working based on the reactive mode when Utilization/Capacity KPI exceed a pre-defined threshold.

Capacity Planning is an AI-powered solution that helps the radio engineers in the process of obtaining optimized network dimensioning to achieve most efficient use of the capacity, for current and future traffic. Capacity Planning support LTE/NR network technologies. The AI-enabled

prediction engine includes a traffic forecasting module that provides at cell level/sector level traffic category KPI forecast, allowing dimensioning for future traffic. Capacity Planning also provides optimal dimensioning actions to avoid traffic bottleneck. These actions include traffic balancing, addition of new carriers or sectors, and estimation of new site additions. The time interval for capacity planner could be hourly, daily, or weekly. Nowadays monthly level prediction is rarely used and mainly depends on the long-term monthly level of input raw data.

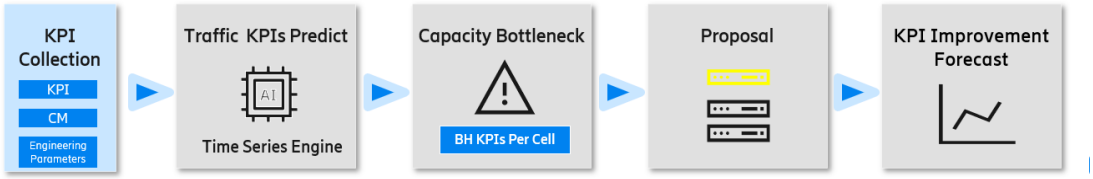


Figure 6-2 Capacity Planner Working Flow

The capacity prediction intelligent solution enables the user to move from a reactive way of planning required capacity to a proactive one based on the combination of traffic forecast and ML utilization KPIs prediction to produce the optimum capacity expenditure per month to meet a certain QoS level.

Main use cases are:

- Forecast predictions
- Identification of bottlenecks
- Dimensioning of the operator's network

6.3.1.3 Application and Performance

Summary of main benefits:

- Shift the dimensioning process from a reactive approach to a proactive one, where bottlenecks can be solved in advance of their occurrence. AI-powered KPI performance prediction model, cell-by-cell for the whole network.
- Identify the right capacity solution to meet the planned QoS goals, eliminating overspend due to inaccurate planning.
- Used for carrier upgrades, new sectors, new sites recommendations.
- Drastic OPEX reduction through automation: 60% of time freed from low value / high manual effort tools.

6.4 Use cases of Network and Service Deployment

There still seldom applications about network and service deployment, and the relevant use cases are still an open issue and topic for all partners to contribute.

6.5 Use Cases of Network and Service Maintenance

6.5.1 Energy Saving

6.5.1.1 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.

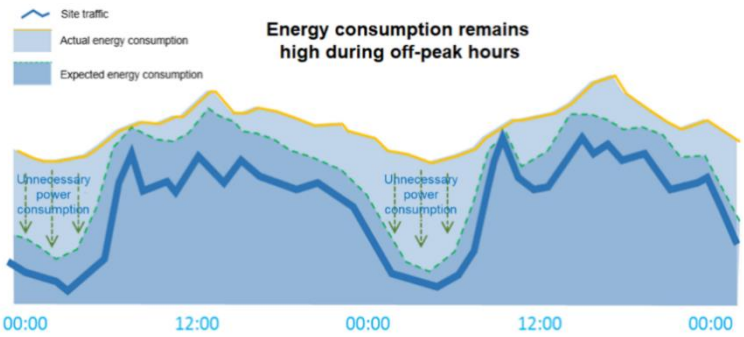


Figure 6-3 Challenges facing traditional energy saving

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.

6.5.1.2 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.

6.5.1.2.1 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.

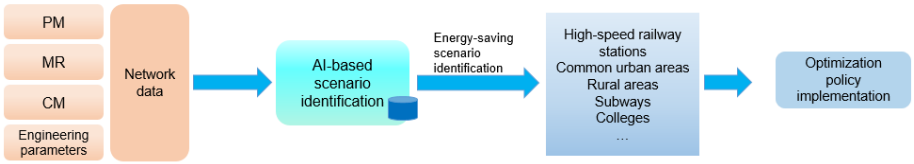


Figure 6-4 Scenario identification for energy saving

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.

6.5.1.2.2 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.[59]

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.

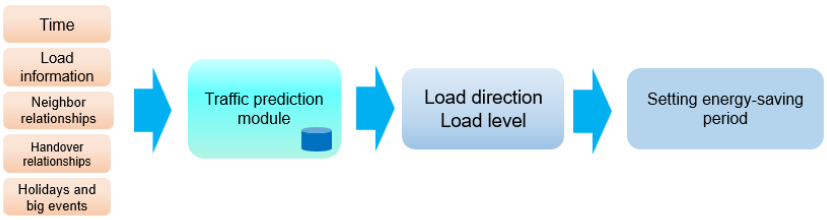


Figure 6-5 Traffic prediction for energy saving

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.

6.5.1.2.3 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.[60]

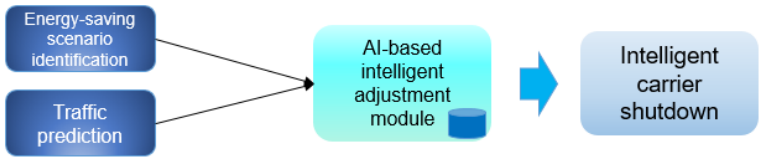


Figure 6-6 Intelligent carrier shutdown for energy saving

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.

6.5.1.2.4 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,

$$\alpha \times (1 - \text{call drop rate}) + \beta \times \text{access success rate} + \gamma \times \text{normalized average throughput} + \mu \times \text{normalized energy consumption}$$

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 6-7 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.

6.5.1.2.5 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.

6.5.1.3 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.

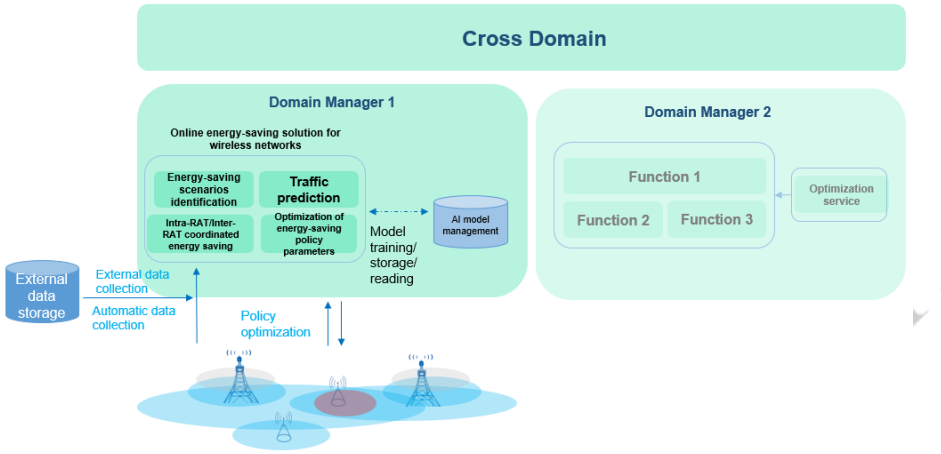


Figure 6-8 Architecture of the online energy-saving solution for wireless networks

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%.

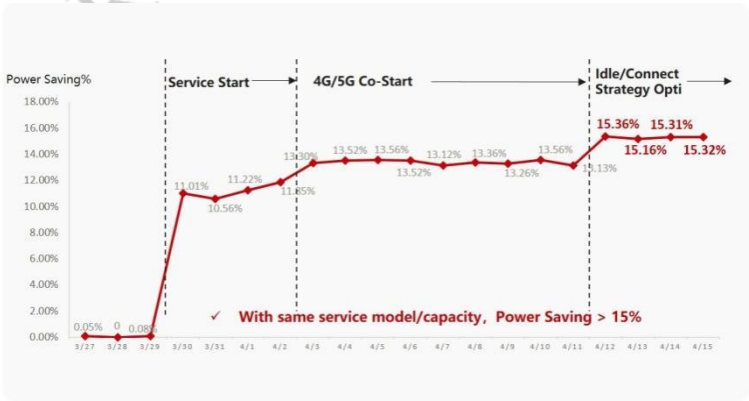


Figure 6-9 Simulation result of online energy-saving solution

6.5.2 AI Powered MIMO Sleep

6.5.2.1 Background

MIMO sleep as a basic power saving function, reduces the power consumption with switching off some antennas when there is no traffic or low traffic. MIMO sleep is widely used in 4G and it is very important power saving function in 5G as 5G products might consume much more power than 4G, especially for AAS. The basic Massive MIMO sleep only enables the function during a period with the threshold. To get more power saving, machine learning is a technique that analysis the data, finds patterns in data, and makes predictions on traffic, and then switch off/on antenna to save more power consumption.

6.5.2.2 Solution Overview

The traditional solution of MIMO sleep will cause the performance degradation due to some transient time. For this transient time, sometimes it might not be suitable for enabling MIMO sleep.

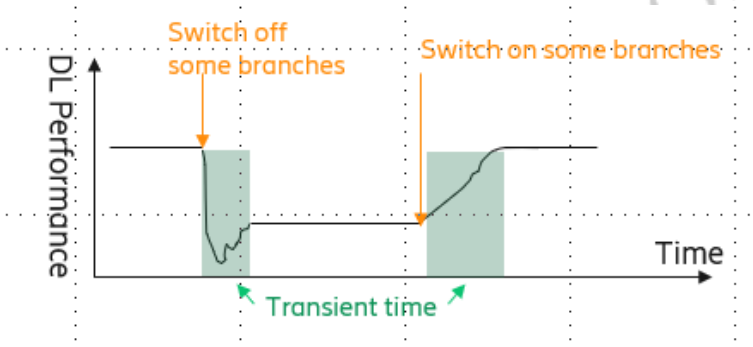


Figure 6-10 MIMO sleep transient time

With AI enabled MIMO sleep, collect the data and analysis the data, build a machine learning model such that an algorithm can use it to make predictions for future traffic.

With the real data, it can easily divide the scenarios into three scenarios: perfect energy saving area, energy saving with capacity degrade, not suitable for power saving. Different scenarios need to be different strategy. In perfect energy saving area, can switch off more antenna to save more power. For some areas, can't enable the function, otherwise will cause capacity and performance degradation or impact the coverage.

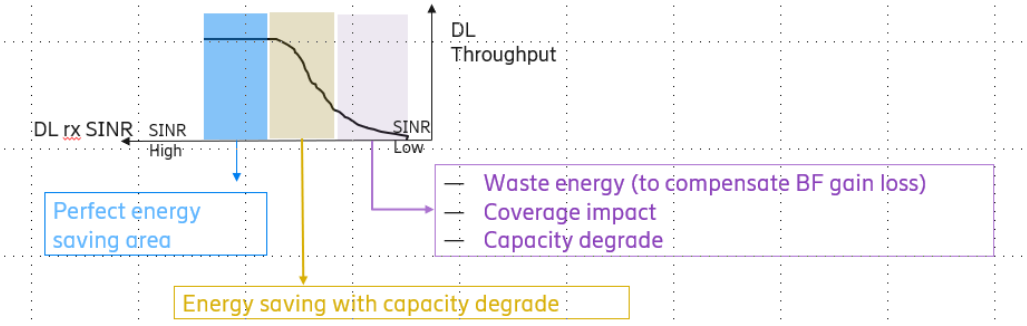


Figure 6-11 Three scenarios with MIMO sleep energy saving

Based on the traffic prediction foe these three scenarios,

- MIMO sleep will be enabled or disabled based on the prediction result.
- The MIMO sleep parameters can be modified as well.

It can find more suitable time to switch off antenna to save power consumption based on the AI enabled.

6.5.2.3 Application and Performance

With ML techniques that are used for traffic conditions forecasting, AI powered MIMO sleep mode can reduce power consumption in the range of 8~14%.

Reduce OPEX by full automation, i.e. no need for manual configuration of MIMO switching criteria for each cell individually.

6.5.3 Root Cause Analysis of Alarm

6.5.3.1 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.

6.5.3.2 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.

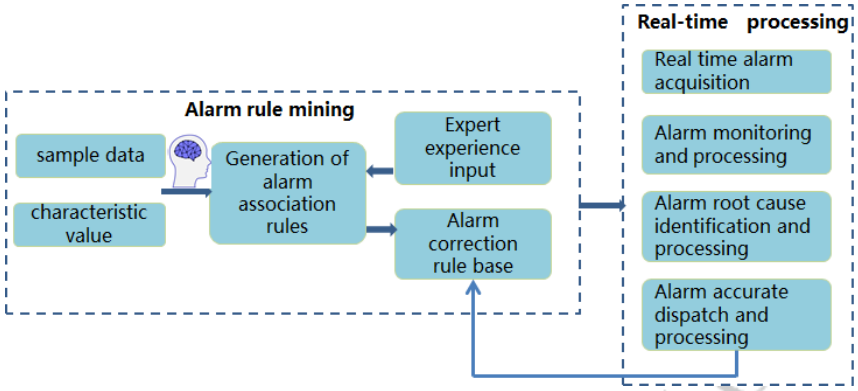


Figure 6-12 Solution chart

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. 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.

6.5.3.3 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%.

6.5.4 Root Cause Analysis of Cell Performance Issue

6.5.4.1 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 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. Network engineers can handle increasingly complex network problems.

6.5.4.2 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 massive 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.

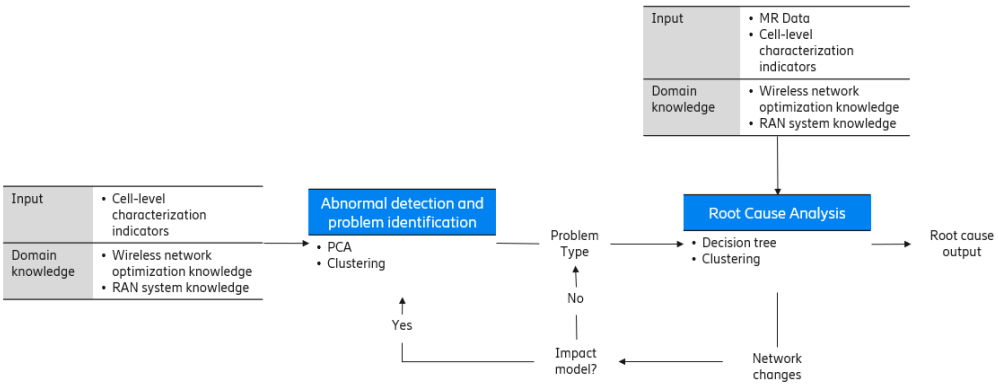


Figure 6-13 Root cause analysis system flow chart

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.

6.5.4.3 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 > 88%

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.

6.6 Use Case of Network and Service Optimization

6.6.1 Autonomous Neighbor Cell Configuration and Optimization

6.6.1.1 Background

Neighbor cell configuration and optimization is one of the most important work for Cellular network planning and operation. Because neighbor cell configuration can influence the network

coverage and user interworking performance directly. Previously, the configuration of neighbor cell relationship mainly relied on expert experience or automated judgment tools based on self-organizing networks. The former mainly adds and optimizes the cell neighbor relationship by analyzing test data, background performance data and geographical environment analysis. It is inevitable that there will be some problems such as neighbor cell missing, one-direction configuration, and incorrect configuration of neighbor cell relationship. The latter has a mismatch rate of more than 30%, and lacks of further optimization for the mismatch and redundancy of neighbor cell relationship, so it increases neighbor cell configuration efficiency at the expenses of user's interworking experience.

With the construction and popularization of 5g network, the importance and difficulty of 5G network related optimization have long been highlighted. While accumulating 5G network optimization experience, operators urgently need intelligent tools to provide support for 5G network planning, optimization and operation and maintenance.

Therefore, operators face two major challenges in the optimization of neighboring areas:

- Due to the limitations of traditional optimization methods, it needs to continuously invest a lot of human and material resources from the cell neighbor relationship configuration planning in the early stage of network construction to the cell neighbor relationship optimization during the daily maintenance in the middle and later stage of network construction.
- The existing automatic cell neighbor relationship configuration tools generally have the problem that the neighbor relationship is easy to add and difficult to delete, resulting in an increasingly large cell neighbor relationship list and affecting the user experience. Relying on expert experience to delete the redundant neighbor relationship, there are risks such as low accuracy of adjacent cell relationship identification, multiple deletion, wrong deletion and so on.

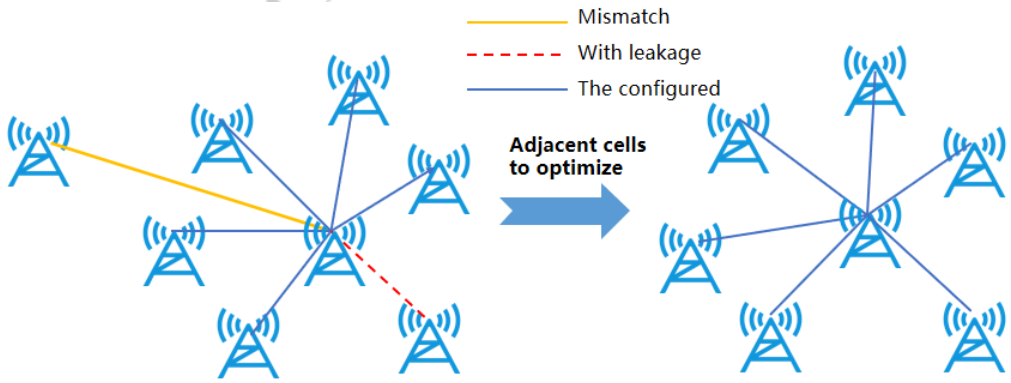


Figure 6-14 Schematic diagram of neighbouring cell relations

Based on the above background, intelligent cell neighbor relationship judgment ability available for 4G and 5G network with high precision and recall is significant for cellular network operators. In this case, an intelligent neighbor cell configuration system is constructed based on cell engineering parameters, user measurement report and inter-cell mobility data. Under network planning scene, only engineering parameters are used for neighbor cell configuration.

6.6.1.2 Solution Overview

Intelligent cell neighbor relationship configuration and optimization provides high-precision neighbor cell list for newly opened cells in the network planning stage through AI algorithm model based on engineering parameters. In network optimization stage, as target cells provide communication service for users and collect user measurement report and mobility information, UE MR and cell interworking performance data are also used for high-precision model building and optimization.

To achieve high-precision neighbor cell configuration, the engineering parameters is used to complete the match on cell pairs to construct the offline training data set. After matching cell pairs, the engineering parameters and inter-cell mobility information are associated. As there is large amount of UE MRs under one cell, there will be a merge and statistic on UE MRs before associated to cell pair. Then, data preprocess comes up with missing value padding and feature engineering. Finally, high-precision neighbor relationship judgment model is constructed through artificial intelligence model design and modeling optimization.

The specific scheme process is described as follows:

- **Training phase:**

Data preprocessing: complete the cell pairs matching through certain rules. Associate the engineering parameters and inter-cell mobility information. Associate UE MRs after data merging and statistic. Finally correct the abnormal data and pad missing data.

Feature engineering: compute cell coverage and co-cover ratio features with engineering parameters and UE MR, generate inter-cell interworking performance features with mobility data and transform the cell's network standard (4G-4G/4G-5G/5G-5G) and their respective coverage scenarios. The last step aims to ensure the model's adaptability to multiple network types and multiple scenarios.

Model training: Based on the two-step combination modeling and optimization of decision tree feature extraction and neighboring judgment, a high-precision model that can be online is formed.

- **Judgment phase:**

Data preprocessing and feature engineering: same as training phase.

Model judgment: use the trained model to complete online real-time inference, and output the conclusion of the relationship between the neighborhood and the neighborhood.

Cell neighbor relationship list sorting and output: Based on application and optimization requirements, sorting and sorting the relationship between multiple cells of the target cell and cell neighbor relationship, and outputting the cell neighbor relationship list.

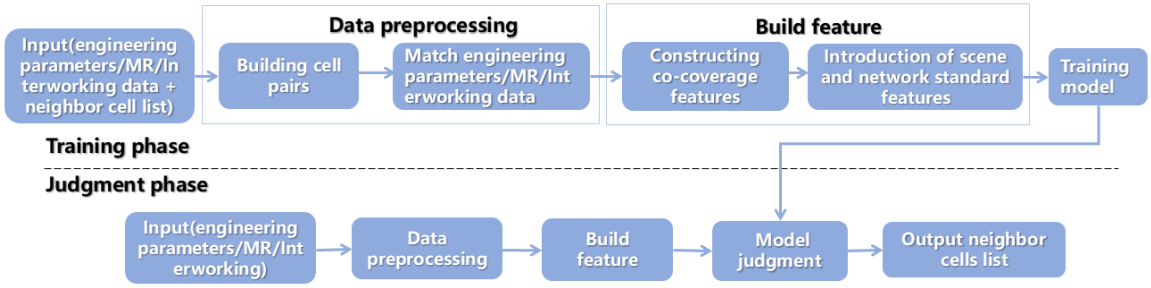


Figure 6-15 Solution for Network optimization stage

6.6.1.3 Application and Performance

In terms of application effect, intelligent cell neighbor relationship configuration system has been deployed in several provinces to provide periodic neighbor cell configuration and optimization service. The configuration accuracy meets the application requirements of the current network.

The results of effect verification are as follows: Compared with the existing tools, the accuracy of cell neighbor relationship configuration is improved by about 30%. At the same time, for the cell neighbor relationship configuration of 5G cell, the accuracy effect exceeds the overall accuracy level. In addition, the product prototype can automatically complete the cell neighbor relationship judgment of new cells / cells with changed work parameters, and output the cell neighbor relationship list of the target cell every day without additional human investment.

6.6.2 Mobility enhancement Based on Machine Learning

6.6.2.1 Background

It is a trend that more and more operators deploy multiple frequencies in their network. Usually, not all frequencies have the same coverage on all places in the network and usually operators would like to configure each base station the same, independent of if there is coverage from other frequencies or base station. Today's implementation of mobility control functionality (similarly for other functionalities as load balancing and s-cell selection) is not optimal in the sense that they use information on what is configured but could not actually present for a specific UE location.

For each UE reaching A2 search there might be a lot of frequencies for the UE to measure on, some of the frequencies might not even be present at that location, that is there are in that case no cells on that frequency deployed that meets the report criteria. The UE will also look for cells on those frequencies, by doing measurements on one of the frequencies (it is not standardized what order of frequencies the UE should measure upon) and if there are cells that meets the measurement criteria, the UE will send a report and HO will be performed. The UE will not look through all frequencies and select the best one, and report that one first. Hence it is very likely that a UE will end up in a suboptimal cell and frequency.

In the example below, when UE moves to a bad coverage area as in the diagram, UE will start A5/A3/B2 measurement for inter or IRAT frequencies. Due to there is not frequency coverage for F2, it is optimal for the UE not to measure F2.

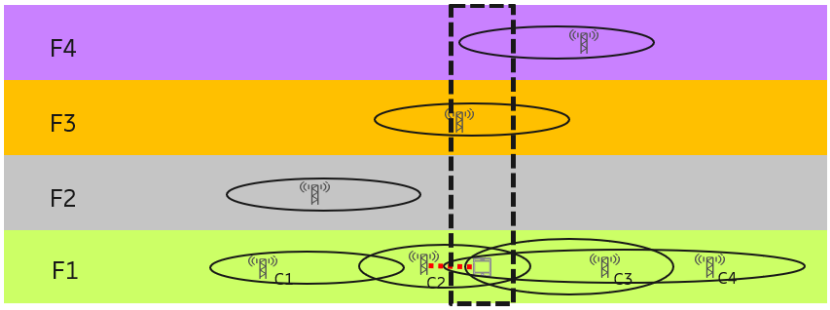


Figure 6-16 UE at poor coverage for inter-frequency or IRAT

Here, the main idea is to enhance the inter-frequency or IRAT measurement procedure by machine learning: learning the correlation between a UE position and existence of other frequency carriers enables the mobility control functionality to make better selection on frequencies for a UE, at the same time, also save the UE measurement time for inter frequency and IRAT.

6.6.2.2 Solution Overview

The solution is to separate the search zone into different areas and based on information from the UE select the best frequency. Example in the green area of the search zone the UE should be selected to measure on F1 not F2 while in the “orange” area the UE can be moved to F2.

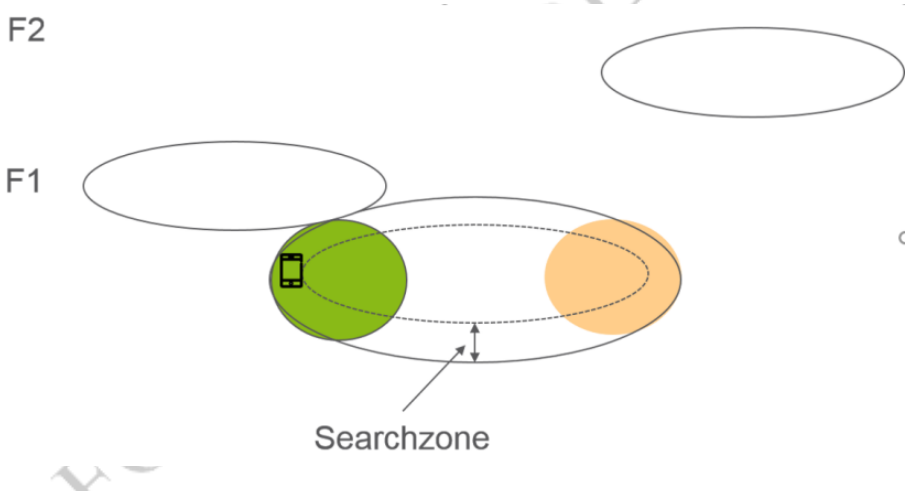


Figure 6-17 UE select different carriers to measure based coverage location

Main solution idea is to introduce machine learning in order tom learn the correlation between a UE location and what frequencies that exists on that location. Here the location will be a RSRP vector (measured cells on the same frequency).

Figure 6-18 Input and output of ML model

Machine learning as such is then divided into a number of different parts: Training, Execution and

re-Training phase

Training phase

Where you can say the correlation between a selected input X and output Y is learned. Simplified, this could be seen as learning a function g, that for every possible input of X finds the output Y. Normally this function is referred to as a model

Execution phase

Where the input X, in this case a measurement report on serving frequency, is used to estimate expected RSRP on other frequencies by using Y = g(X)

Re-Training phase

Where one discovers that based on measured input X and measured output Ý deviates from the calculated output Y = g(X), then retraining is needed.

6.6.2.3 Application and Performance

With ML model to predict candidate carriers for UEs to avoid unnecessary measurement for some carriers, inter-frequency handover measurement time can be reduced by 50% in some field tests, as indicated in Figure 6-19.

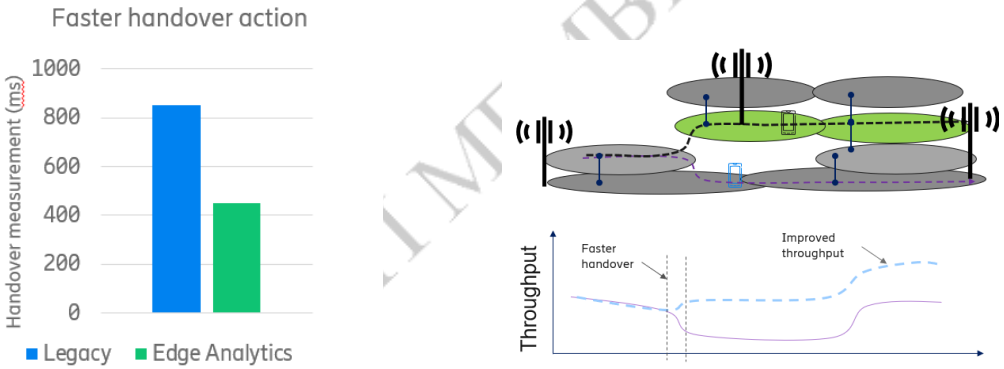


Figure 6-19 Handover time save and UE throughput enhancement during handover

And due to UE will stay in poor coverage area for a shorter timer, UE average throughput will also be improved, as indicated in Figure 6-19.

6.6.3 Abnormal Detection Trial

6.6.3.1 Background

Since telecommunication networks are becoming more and more complex, continuous network monitoring is essential in detecting and reporting failures as soon as possible. In the current network operation, the detection and handling of abnormal/problem cells still highly depend on expert's experience and network operation effort is high. As shown in Figure 1, in the traditional solution, the detection of abnormal cells is carried out based on the 'hard' statistical threshold, which is defined according to the experience. The shortcomings of traditional methods are as

follows:

- New anomalies cannot be detected. Traditional methods do not have corresponding threshold / index to detect new anomalies.
- The anomaly indication is not accurate enough by using the traditional methods, the cell anomaly detection threshold calculated by limited data/indicators. Thus, the characteristics of anomalies lack fine description and portrait, The same anomaly often contains anomalies with completely different causes.

Above two problems will potentially block network operation automation and increase the operation effort.

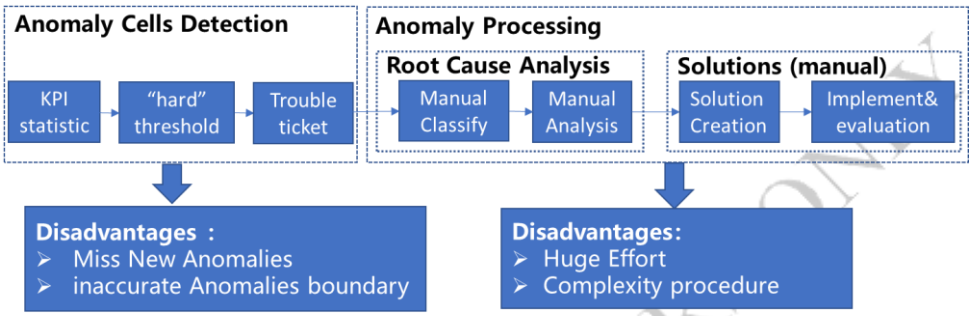


Figure 6-20 Traditional abnormal/problem cells detection procedure

Aiming at the above problems of traditional abnormal cells detection procedure, a multi-dimensional time series anomaly detection scheme , which is based on AI / Ml solution, is proposed to realize real-time anomaly detection for the huge KQI/KPI data of the network, and realize the following functions:

- For anomaly detection: unsupervised anomaly detection detects new anomalies in real time, which can be used as a supplement to the existing detection rules.
- For analysis: intelligent root cause analysis to reduce human effort

6.6.3.2 Solution overview

As shown in Figure 6-21, the anomaly detection system mainly contains two functional modules: one is for model training, the other is for detection. The model training module collects the KPI/KQI data from the cell and perform model training in period. From the figure, the model was trained using a multi-dimensional anomaly detection algorithm, which analyze and cluster hundreds dimensional KPI/KQI indications/data time series. The several anomaly classes will be generated automatically according to the KPI/KQI vector's distance. The anomaly class contains the features extracted from hundreds of KPI/KQI indications/data time series, so the anomalies are described more accurate. The detection module detects the network KQI/KPI data in real time, and automatically classifies the detected anomalies according to the anomaly characteristics described by the anomaly detection model.

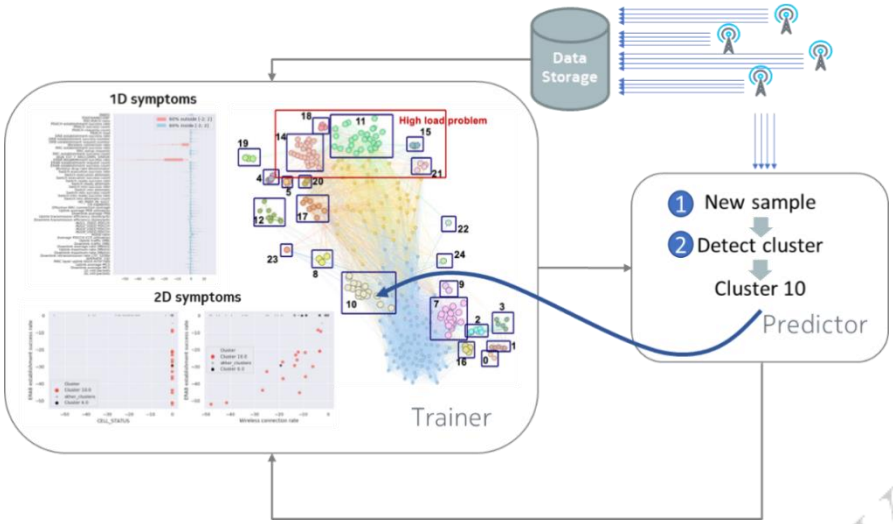


Figure 6-21 Algorithm and architecture

6.6.3.3 Application and Performance

A trial was carried out to utilize Anomaly Detection System in operator's regular radio network maintains procedure and verify Anomaly Detection's function and performance. The trial network is in a province capital city, and the scale of the network is more than 10,000 cells. The Anomaly Detection System detects data including cell KPI (15min) and KQI (5min), total ~300 indicators. The effect of anomaly detection algorithm is evaluated through defined user case and compared with traditional methods.

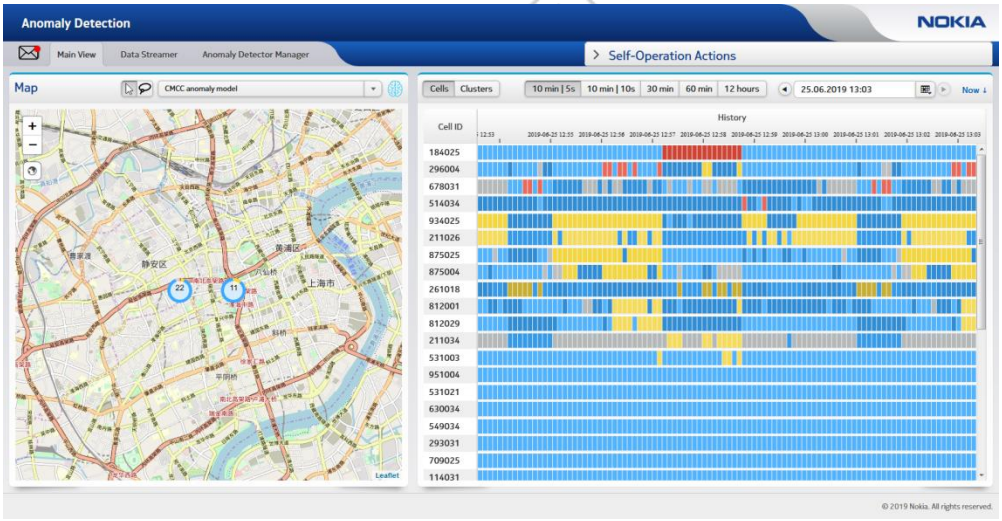


Figure 6-22 Abnormal detection trial networks

Use case: high capacity cell detection

High capacity cell detection is one of the most complex use cases in network maintenance. In the hot traffic area, some cells' capacity is limited by the growth of users and traffic, which usually lead to channel congestions, low user data rate and even user complain as worse user experience. These "true" high capacity cells need to be eliminated through network capacity upgrade. The solutions usually include adding new carriers, deploying new sites etc.... At the same time, some cells also

show high capacity phenomenon but as different causes, such as the capacity limitation and channel congestion due to handover / poor channel quality etc. They are "false" high capacity cells. These "false" high capacity cells need to be eliminated by network optimization. Traditional methods use 1-2 indicators to check high-capacity cells. the detect policy is too coarse to precisely describe the feature of high capacity cell, thus difficult to distinguish between 'true' and 'false' high capacity cells. According to statistic, there are ~20% high capacity cell is 'false' by using traditional solutions. As network upgrade investment is high and construction period is long, thus misjudgment of high-capacity cell will generate great economic losses.

In the trial, multi-dimensional anomaly detection is carried out for more than 300 KQI/KPI indicators. According to anomaly detection mode, ~20 capacity related indicators are found and used to present cell's capacity character. The multi-dimensional KPI / KQI indicators are used to classify high capacity cell. According to the abnormal indicators, the cells high capacity characters are divided into serious, medium and slight. According to the indications/class, operator can easily manage the high capacity cells and make network upgrade plan. The "false" high capacity cell problem is totally eliminated. During the trial, the number of cell capacity ticket, which is warning signal for cell high capacity and need expert's effort to analysis, is reduced by 63%.

UL/DL	Both UL/DL	UL	DL	both
PRB usage high	Yes	Yes	yes	no
User # high	Yes	no	yes	yes
Traffic	Yes	yes	no	no
Packet # high	yes	yes	no	no

Figure 6-23 High capacity cell characters is precisely described by anomaly detection

Table 6-1 shows the processing time of cell capacity ticket in traditional/ manual way and AI / ML anomaly detection way, respectively. According to table 1, the traditional /manual analysis consists of five steps. In each step, different type of data needs to be retrieved and analyzed. The total processing time of each cell capacity ticket takes 30-70 minutes. The AI / Ml anomaly detection method automatically generates the detection/analysis report. The people for network operation process the high capacity ticket according to the suggestions of the detection/analysis report, and the processing time is reduced to 5 minutes. The network operation is greatly improved by AI/ML anomaly detection.

Table 6-1 Trouble Ticket processing effort: Legacy vs. AI

Traditional root cause anaylsis (manual)			AI/ML processing		
	Action	processing time		Action	processing time
step1	collect neighboring cells information of cell report capacity anomaly	5 min	step1	read report generated by AI/ML	5-10min
step2	collect relate KPI/KQI data before and after trouble ticket, analysis the correlation of time sequence.	5-10min			
step3	collect neighboring cells capacity data, analysis the correlation of neighboring cells	5-10 min			
step4	collect network parameters to check handover, power, load balance setting etc..	5-15min			
step5	root cause output	10-30min			
Total		30-70min			5-10min

Summary: compared with the traditional detection methods for high-capacity cells, proposed AI/ML based anomaly detection system precisely describe the high-capacity cells by using multi-dimensional indicators and realize automatic analysis. By eliminating the "false" high capacity cell in capacity problem ticket, the number of capacity problem ticket is reduced by 63%. The processing time is reduced from 30-70 minutes per ticket in the traditional method to 5 minutes. The processing efficiency is greatly improved.

6.6.4 NR Network UE Throughput Optimization

6.6.4.1 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.

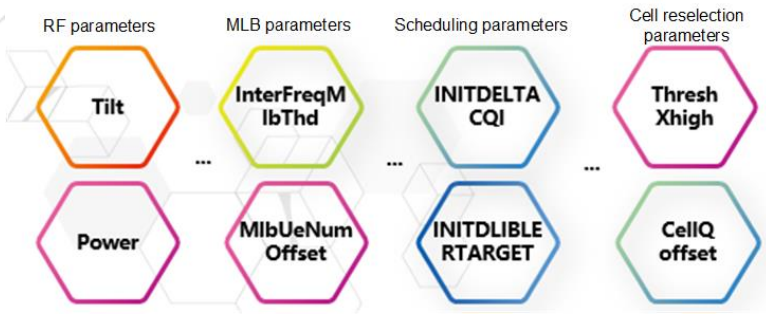


Figure 6-24 Thousands of parameters related to the air interface

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.

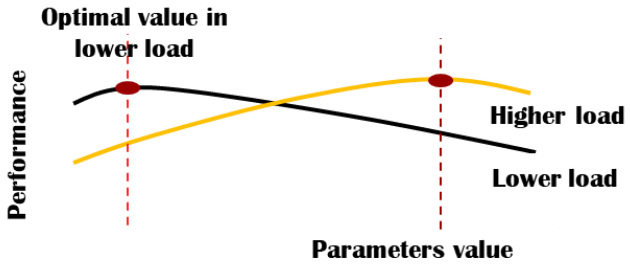


Figure 6-25 Parameter settings varies in different scenarios

6.6.4.2 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.

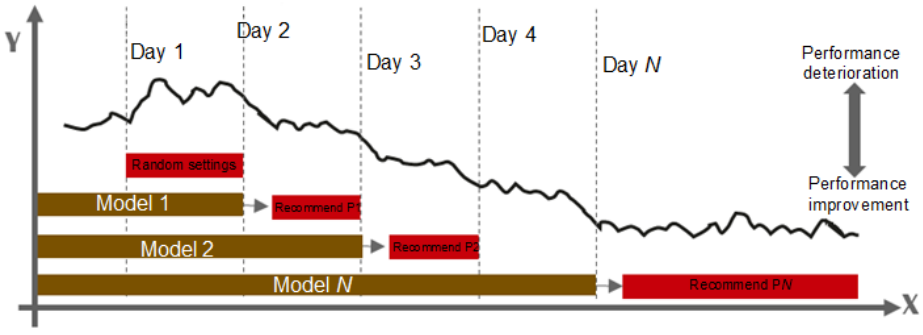


Figure 6-26 Iterative optimizing the model

- 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.

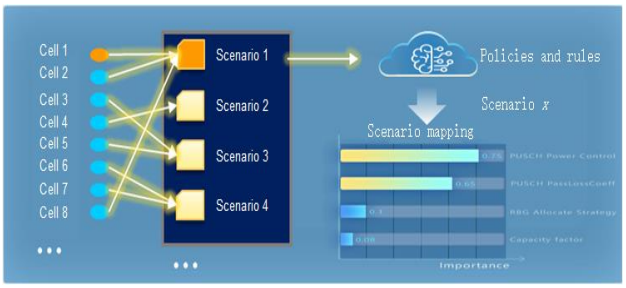


Figure 6-27 Mapping different scenario with proper parameters

1.1.1.1. 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%.

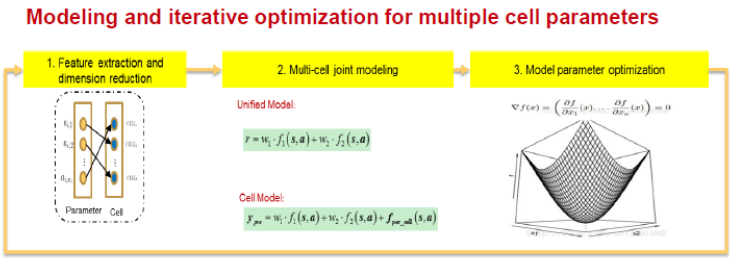


Figure 6-28 Modelling and iterative optimization for multiple cell parameters

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.

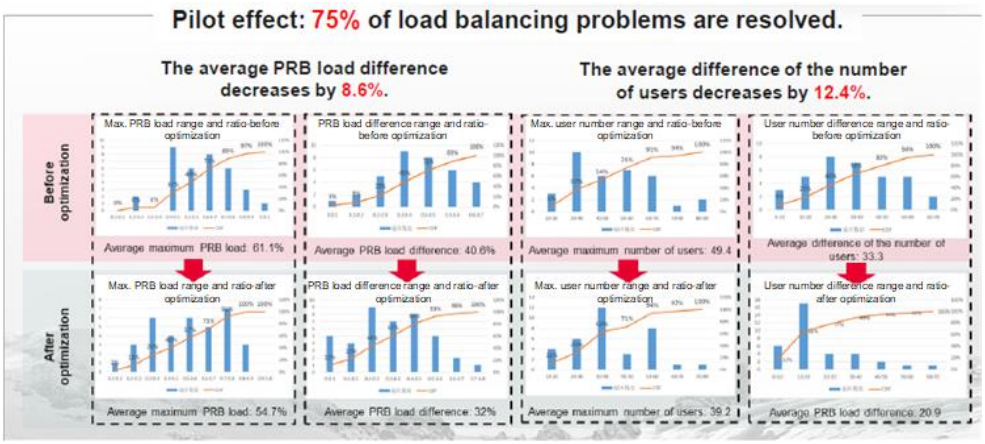


Figure 6-29 Pilot result of load balance with intelligent multiple parameter optimization

6.6.5 NR Network Coverage Optimization

6.6.5.1 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.

6.6.5.2 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.

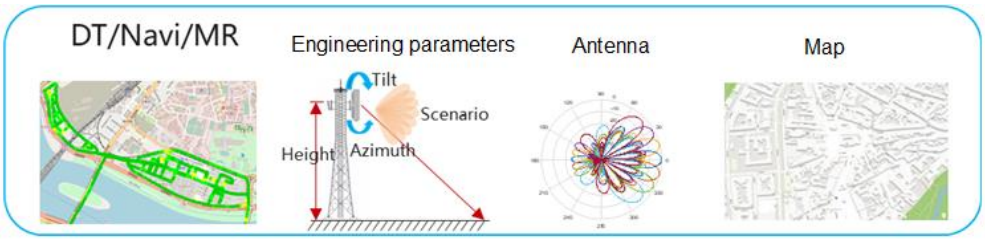


Figure 6-30 Multiple dimension of data collection

- 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.

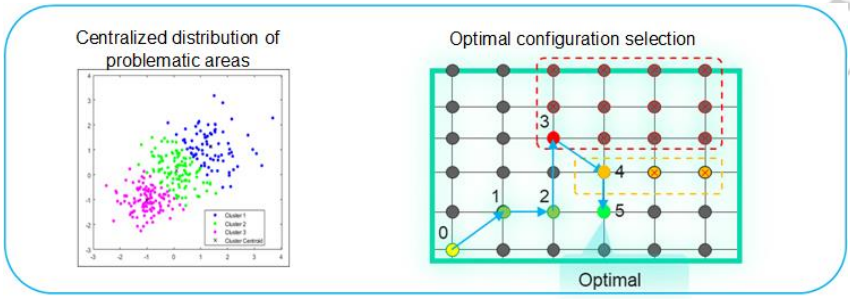


Figure 6-31 Locate the problematic area and find the optimal configuration

- 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.

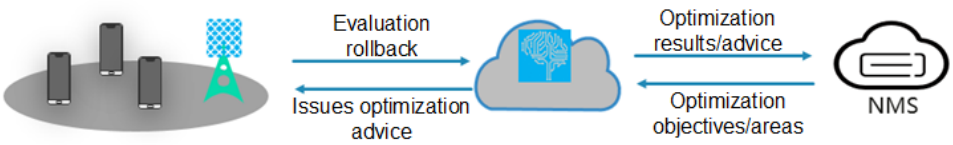


Figure 6-32 Evaluate and verify the optimization advice

6.6.5.3 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.

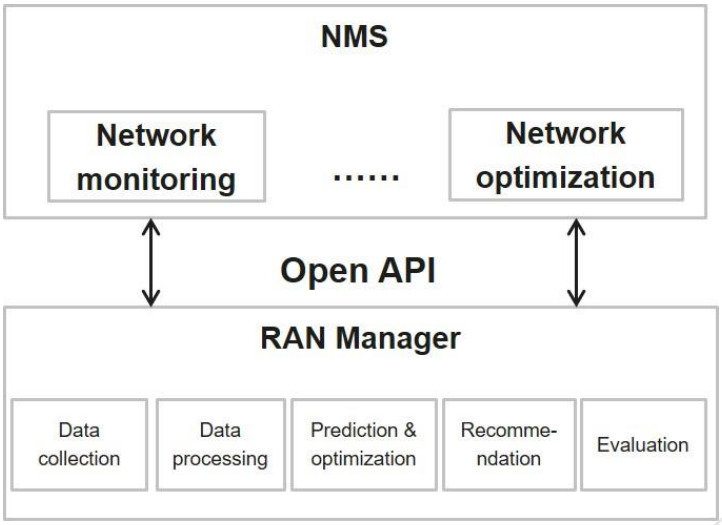


Figure 6-33 Network framework in a typical operator application scenario

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.

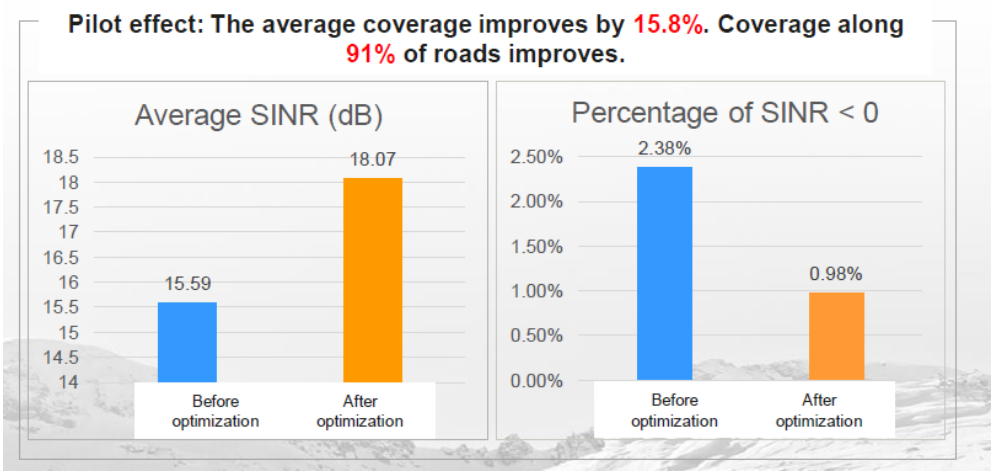


Figure 6-34 Pilot result of NR network coverage optimization

6.6.6 ML-Based MU-MIMO Scheduler

6.6.6.1 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 figuring 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.

6.6.6.2 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.

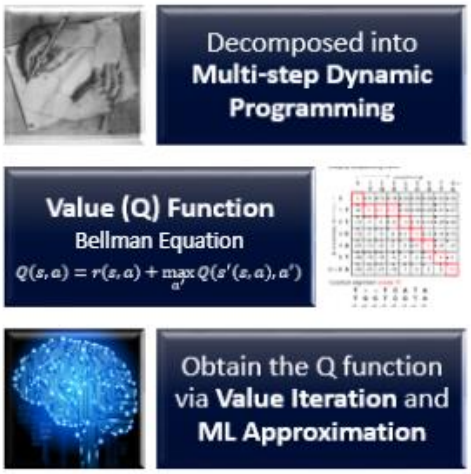


Figure 6-35 Deep Q-learning or reinforcement learning

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.

6.6.6.3 Application and Performance

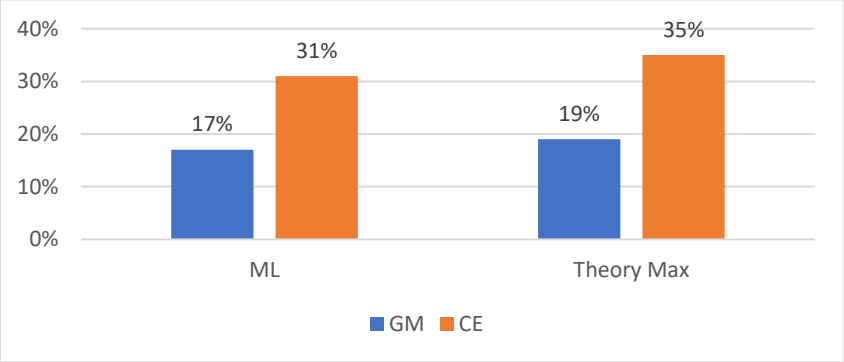


Figure 6-36 Simulation results of ML-based MU-MIMO scheduler

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.

6.6.7 Link Adaptation

6.6.7.1 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.

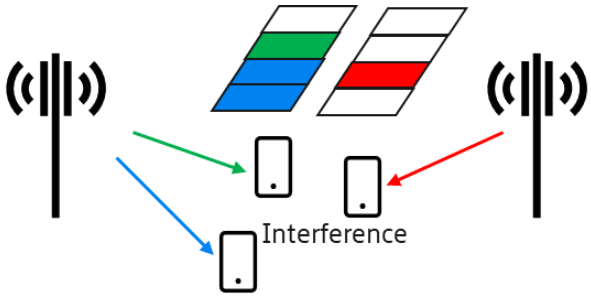


Figure 6-37 Interference created by burst traffic

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.

6.6.7.2 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.

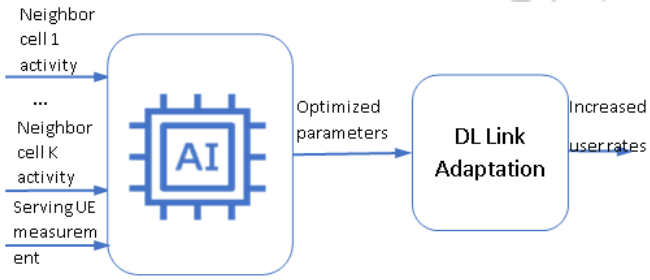


Figure 6-38 Intelligent DL Link Adaptation

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 tunes 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 homogeneous parameters setting.

6.6.7.3 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.

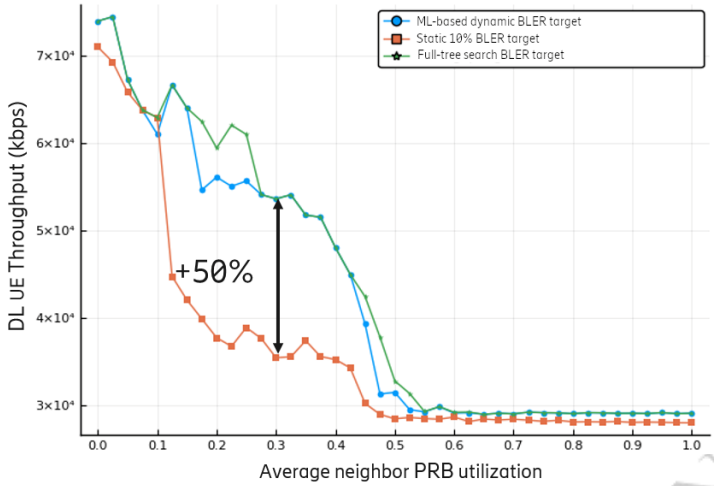


Figure 6-39 Simulation result of intelligent link adaption

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.

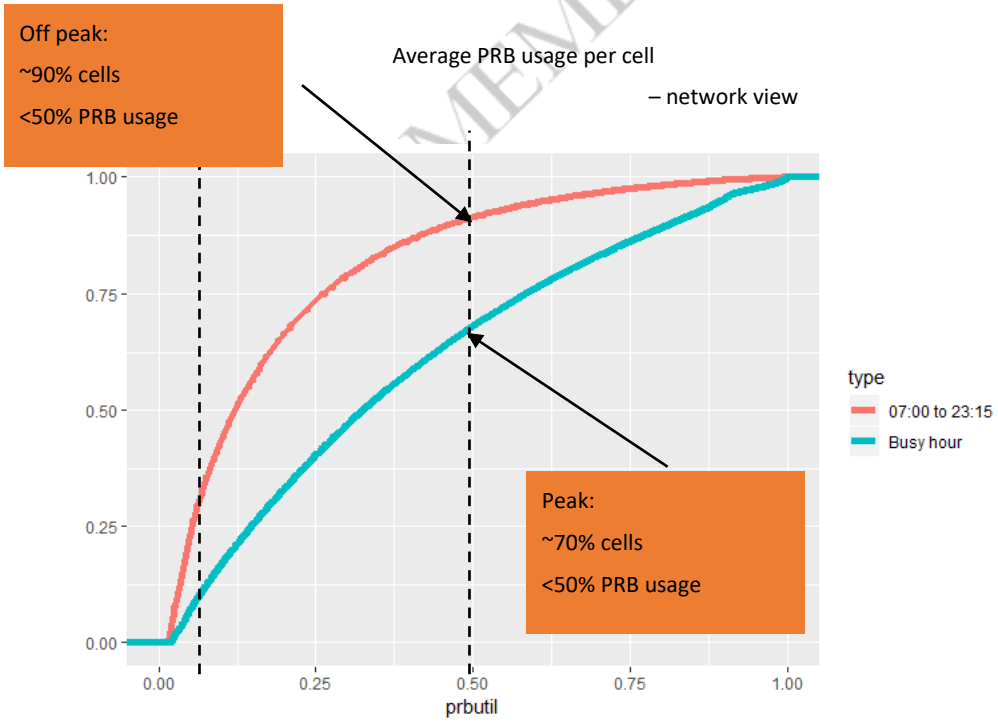


Figure 6-40 Simulation result of intelligent link adaption

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.

6.6.8 Load Balancing Based on the Virtual Grid Technology

6.6.8.1 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.

6.6.8.2 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.

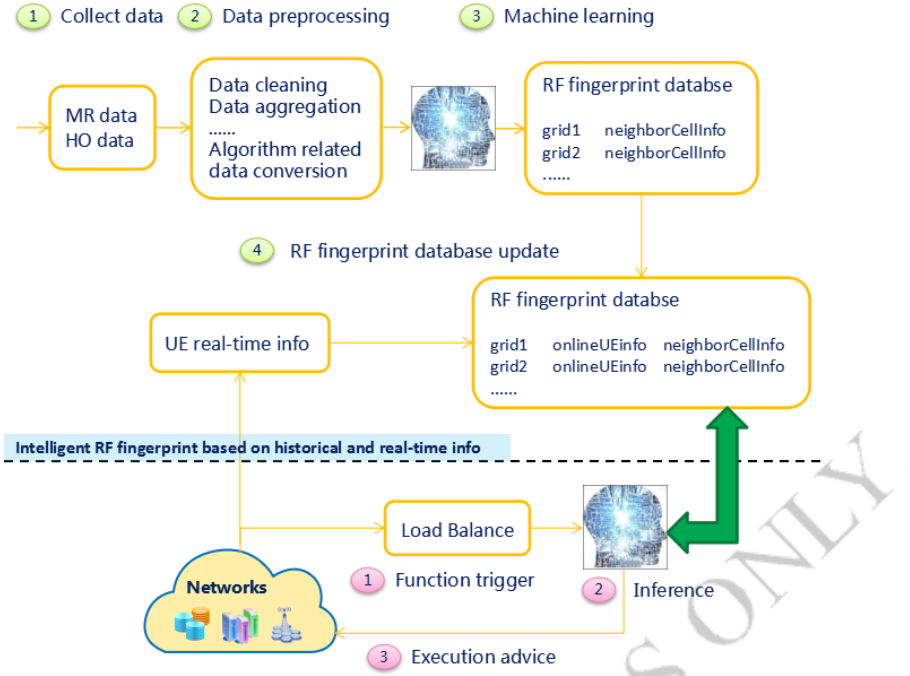


Figure 6-41 Basic flow chart

6.6.8.3 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 6-2 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.

6.6.9 QoE Optimization

6.6.9.1 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.

6.6.9.2 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.

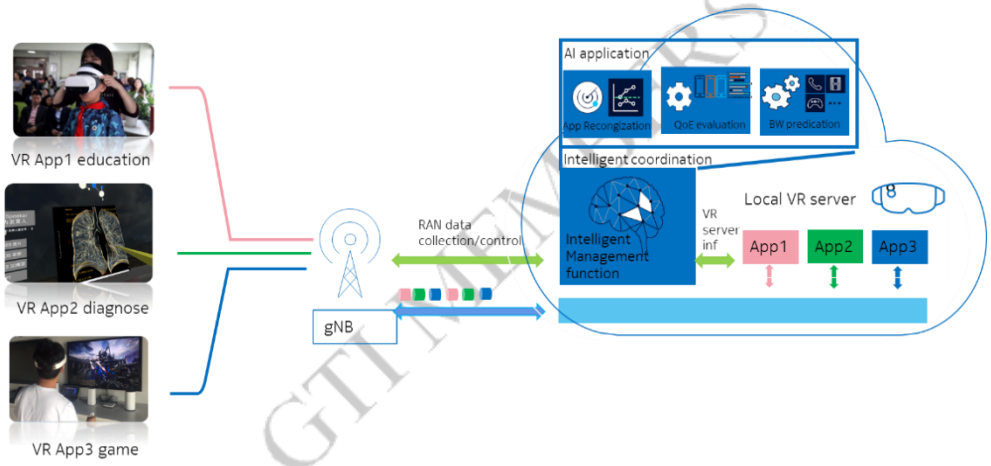


Figure 6-42 Intelligent management function used in VR APP for QoE optimization (example)

6.6.9.3 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 6-43 AI powered Cloud VR experience trial on China Mobile live 5G network

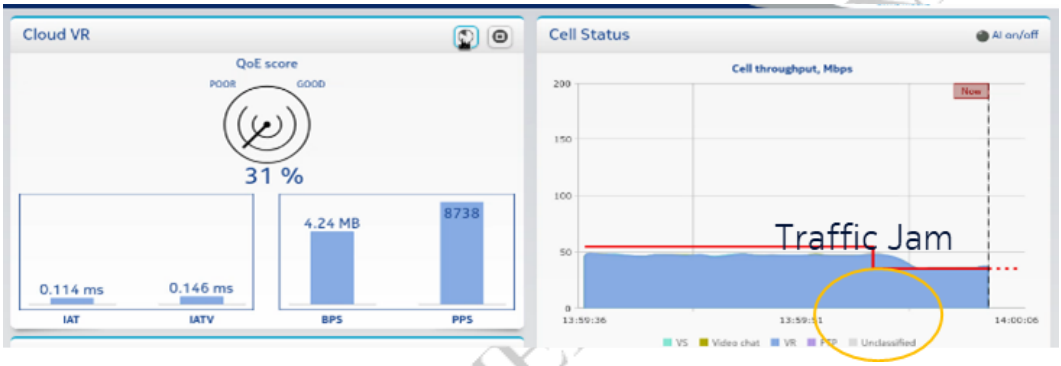


Figure 6-44 trial results: VR game QoE evaluation without QoE optimization

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

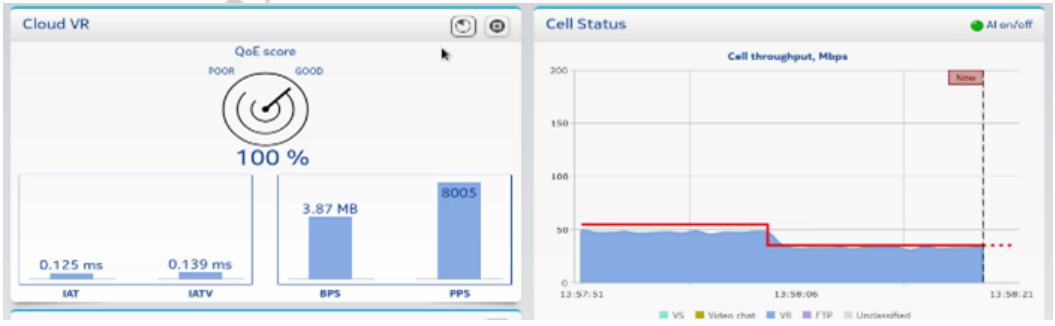


Figure 6-45 trial results: VR game QoE score with QoE optimization

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).

6.6.10 Edge QoS

6.6.10.1 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.

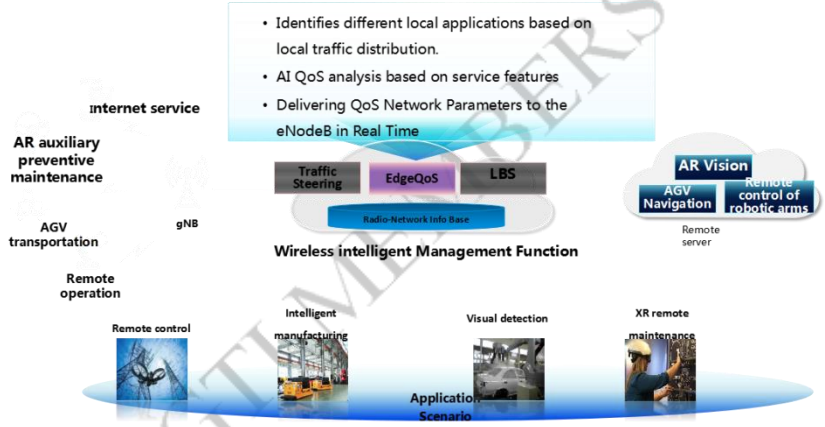


Figure 6-46 Edge QoS application scenario

6.6.10.2 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.

6.6.10.3 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.

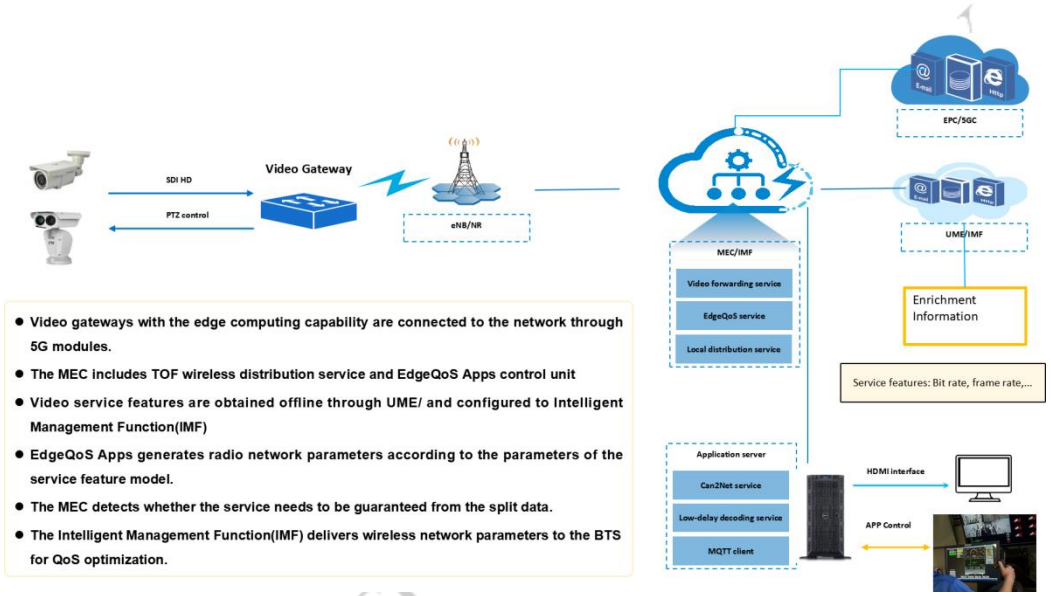


Figure 6-47 Edge QoS remote control guarantee application

6.6.11 Transport Network Optimization

6.6.11.1 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.

6.6.11.2 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.) is 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.

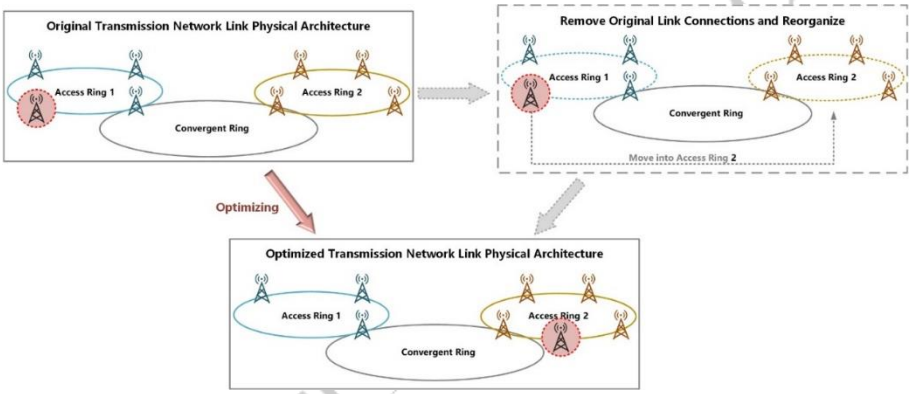


Figure 6-48 Original and Optimized Transport Network Link Physical Architecture

6.6.11.3 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.

6.7 Use Case of Network and Service Operation

6.7.1 Subscriber Complaint Handling

6.7.1.1 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 takes 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.

6.7.1.2 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.

6.7.1.3 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.

6.7.2 Intelligent Video Service Quality Analysis

6.7.2.1 Background

With the development of 5G network technology, how to perceive and guarantee the user experience in terms of video service has become an important means for operators to improve their competitiveness.

Operators' assessment of network quality is shifting from network performance to service experience, but there is currently a lack of efficient and accurate means for end-to-end quality assessment of video services. Mainly reflected in the following aspects:

- End to end video quality analysis needs to obtain user-side experience data, which is collected and privately owned by OTT manufacturers. As a service provider instead of content provider, mobile net operators can hardly obtain user experience information
- The traditional dial test method for video perception evaluation will consume a lot of labor, and cannot achieve comprehensive coverage in space and continuous evaluation in time. Sometime customer complaints are helpful for discovery of pool service quality, but operator cannot achieve global sense for video service quality with them.
- There is plenty video XDR data collected by the DPI system, which aims to describe service quality with information contained in data packets over mobile network. But currently, there is no effective method of video quality analysis with XDR data. At present, the accuracy of evaluating poor video perception solutions based on network-side feature information is between 40% and 60%, and the accuracy of finding problems is low, which is not practical.

In response to these problems, the intelligent video service quality analysis system aims to achieve accurate and efficient service quality evaluation with video service feature data collected by DPI system from network. In this case, AI algorithm and rule base are combined in offline training model, which can achieve real-time analysis of user perception for all on-demand video services. At the same time, combined with customer generation and operation needs to perform statistical summary, automatically output overall business operation information, poor quality user and cell information, and provide poor quality geographic location distribution information, helping front-line service guarantee personnel to quickly understand user-level and overall service quality information, and improve Work efficiency.

6.7.2.2 Solution Overview

The intelligent video service quality analysis system is composed of offline training module and online analysis module. Offline training part is based on data set formed by the association of video service XDR data on network side and the video service experience on user side, and the designing and tuning of artificial intelligence model. The high-precision and accurate service quality evaluation model is applied in online analysis to work out user side experience for video service in mobile network. Finally, the service quality information is presented through the front-end interface, and the data statistical analysis is performed through the back-end analysis system to support and guarantee the service quality of cellular network.

Offline training: To collect user experience information, offline training system collect training data with dial test tool and automatic dial test script. After combining user experience information and user-plane video service information gathered from DPI system deployed on network side, high-quality user experience evaluation training data is collected. Then the evaluation model is formed through many efforts on feature engineering, algorithm choosing, optimizing and hyper parameter tuning. The feature engineering includes feature importance analysis, feature selection and feature generation. Feature importance analysis is enhanced through XGBoost algorithm and business knowledge. Feature selection is performed through feature mutual information analysis and importance ranking through MIC+RFFECV. Feature generation is designed to extend information input to model. The service quality evaluation model consists of multiple base model including decision tree, neural network, recommendation model and SVM, with a stacking structure. Besides, GAN is used to generate and retrain on negative samples which is poor quality data in this case, to improve the analysis ability for poor service quality, taking into account the model The overall accuracy and business quality analysis capabilities of the company have reached an evaluation accuracy rate of 91%, a recall rate of 83%, and an overall accuracy of 87%.

Online data analysis: In this part evaluation models generated from offline training module are used in service quality analysis for video service XDR session data generated from cellular network user. The service quality information is presented through the front-end interface, and the data statistical analysis is performed through the back-end analysis system to support and guarantee production work.

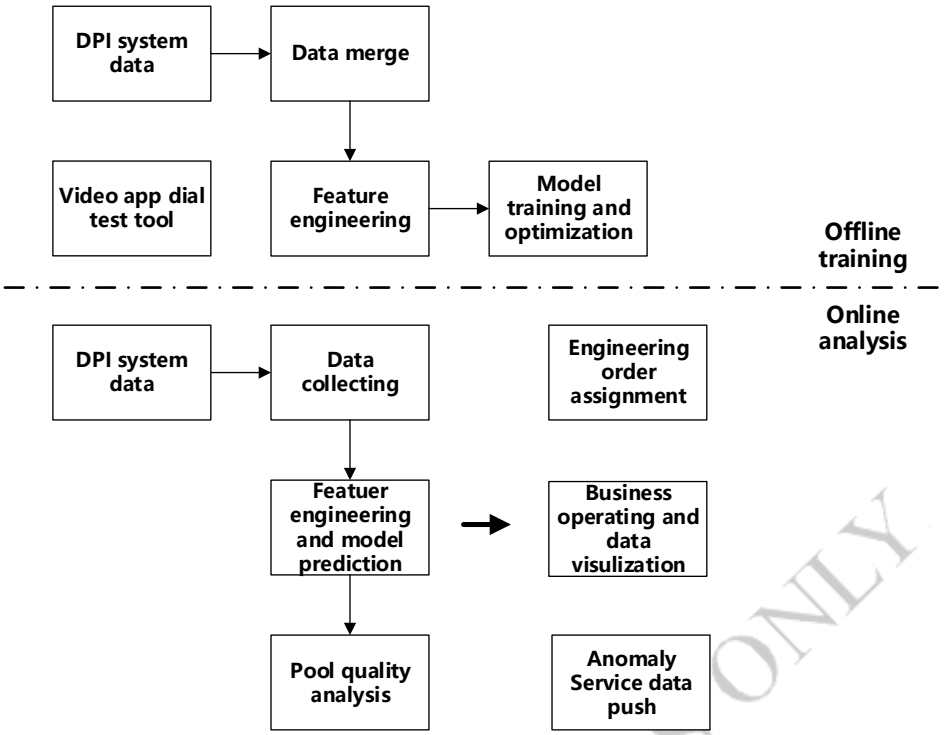


Figure 6-49 Basic flow chart

6.7.2.3 Application and Performance

The intelligent video service quality analysis system is deployed on big data analysis platform in several provinces, and provides service for average daily 2 billion pieces of video service data with quasi-real time service analysis. With the application of intelligent quality analysis system, efficiency of poor quality service detection is highly improved. It only spends several minutes rather than one or two days to discover most currently occurred pool quality service. On the other hand, intelligent video service quality analysis uses AI algorithms to improve data feature quality, and uses a hybrid model composed of multiple basic models to increase the accuracy of service quality assessment to more than 90%.

The intelligent video service quality analysis system provides front-end web pages for efficient network optimization and pool quality backtracking. In addition to scale service quality analysis, the system also provides automatic user-level and cell-level quality analysis and result output.

6.7.3 Automatic Wireless Broadband Service Provisioning

6.7.3.1 Background

Because of the dynamic wireless environment, conventionally, it's difficult to accurately predict which areas can develop FWA service and the throughput it can provide. The traditional way to do it is for the operators to implement drive test in the whole network (pre-sale coverage or capacity map) or even send engineers on site to do signal test every time they need to install the FWA CPE in the customers' premises. It is a very time and human cost consuming process and above all, the accuracy is low due to the manual process.

6.7.3.2 Solution Overview

Based on AI technology, automatic FWA service provisioning solution is able to provide highly reliable pre-sale evaluation in terms of the location, package and CPE it can provide to customers based on coverage, throughput and resources of the network. By connecting with operators' Business and Operation Support System (BOSS), what the operator need to do is just input the name of the target address in the search box, and the pre-sale evaluation result can be show automatically for customer to make the decision, including whether or not the address is able to provide FWA service, what kinds of data package is available and what kinds of CPE are recommended.

AI plays a pivotal role in the automatic FWA service provisioning solution in both pre and post-sale stages.

In the pre-sale stage, in order to provide fast service provisioning, a precious FWA service provisioning map is necessary based on cell capacity and virtual grid level throughput.

First, customers setup the data package and related CPE types. Secondly, by simulation and collecting various data from the wireless network, the built-in AI engine use AI algorithm to build grid level coverage & spectral efficiency characteristics data base and implement the training of AI model, which includes the correlation between user level characteristics, cell level characteristics and ML training target.

Thirdly, when a customers' home address is searched and identified on the provisioning map, the grid level characteristics are extracted and make the query from the spectral efficiency characteristic database built by AI and the grid level spectral efficiency can be predicted. Then, by obtaining the remaining RB resource, the grid level throughput capability can be calculated and the most suitable CPE can be selected automatically.

In the post-sale stage, the suspected churn user can be identified with AI technologies.

First, history FWA CPE log is extracted and if the CPE is not connected in a certain period of time, then it will be regarded as churn user.

Secondly, church user model can be built with AI algorithm by obtaining some characteristic from churn user CPE log and make the correlation with the churn users.
Thirdly, once the churn user model is built, the suspected churn user can be identified by querying the characteristics from the AI churn user model.

6.7.3.3 Application and Performance

Traditionally, it takes at least 51 hours for the customer to know whether their applied residence is serviceable, and customer only brings home a piece of paper as proof of purchase until service is validated through truck roll. Compared with the uncertainty and suffering before, now, front-liners in the business hall of the operator G in South East Asia can inform the customers within an hour about the Plan options with 16% increase of the accuracy of service provisioning.
Besides, with the support of pre-sale evaluation, the engineers don't need to travel on-site to customers' premises for coverage/throughput test and installation, which dramatically reduces the burden of travelling in the extreme weather, especially in South East Asia countries like Philippine, where the yearly average temperature is above 30°C.

7 Essential Capacities

7.1 Introduction

As the autonomous networks applied in data identification & assessment, fault location, alarm & prediction, configuration optimization and other application scenarios, all of the scenarios have been refined into four types of four essential capacities: autonomous perception, autonomous diagnosis, autonomous prediction and autonomous control.

- Autonomous Perception: the capacities which include network and service data collection and necessary data pre-processing with the purpose of monitoring network and service information.
- Autonomous Diagnosis: the capacities which evaluate and decide the necessary operation for execution, e.g. network configuration or adjustment.
- Autonomous Prediction: the capacities which predict the obtained network and service information or based on the historical network and service information to further predict the future change trend of the above network and service status, and make recommendation for decision.
- Autonomous Control: the capacities which execute the operations.

7.2 Autonomous Perception

7.2.1 Description

Network perception is the basic prerequisite for closed loop network services and automatic driving. Therefore, devices must have sensing components are introduced, and the capability of sensing resources, services, and surrounding environments are increasingly stronger. Multi-dimensional real-time sensing is provided, covering service flows, resources, topology status, O&M events, and power consumption.

As major Network Autonomous Perception function, Situation Perception is defined as "The perception of data and behavior that pertain to the relevant circumstances and/or conditions of a system or process ("the situation"), the comprehension of the meaning and significance of these data and behaviors, and how processes, actions, and new situations inferred from these data and processes are likely to evolve in the near future to enable more accurate and fruitful decision-making" [58].

It consists of five actions: gathering data (perception), understanding the significance of the gathered data (through both facts and inferences), determining what to do (if anything) in response to a given event, making a decision (or set of decisions), and performing those actions. It enables the application of context and policies to a particular situation and can use inference as well as historical data to understand what is happening at a particular context, why, and what (if anything)

should be done in response.

7.2.2 Implementation

When separate from decision-making and action specification functions, the Network Autonomous perception including flowing 3 function blocks

- **Perception**: produces a perception of situational elements (e.g. objects, events, people, systems, and environmental factors) and their current states (e.g. modes and locations).
- **Comprehension**: examines the situational elements that have been perceived in order to better understand how they fit together; this helps characterize the situation as a whole, and how this situation affects the goals.
- **Projection**: Network Autonomous perception is usually the basis of prediction and work together to predict the most likely evolution of the current situation.

Network Autonomous perception is fundamental function of an AI/ML user case, and different implementation/solution will be used according to the objects, events, environments of the user cases. Following are some examples of Network Autonomous perception implementation/solution.

1, Application perception

Create real time cross-layer data analysis to generate actionable insights, automatically categorize application flows based upon traffic signatures, apply policies in the RAN based upon real time insights. Optimize spectrum efficiency (PRB utilization) across http adaptive streaming (HAS) clients. Target rates set optimally in line with HAS client UE channel conditions and cell congestion level.

2, Location perception

Accurate user positioning is of paramount importance for industry verticals such as Industry 4.0, where real-time monitoring of assets and robots is critical to the overall business efficiency. User positioning is already included in 5G standards, where it targets meter level accuracies, still far from the cm-level accuracies required in some domains. User positioning is a problem well-suited for the application of AI/ML techniques that can fuse positioning data from different technologies, or aid in determining the line-of-sight path in a multi-path propagation environment.

3, Network Problem perception

As described in Use cases 6.6.3 “Anomaly detection trial”, to quickly detect new anomalies of telecommunication, which became more and more complexity, a multi-dimensional time series anomaly detection scheme is proposed to realize real-time anomaly detection for the huge KQI/KPI data of the network, and realize the following 2 functions:

- For anomaly detection: Unsupervised anomaly detection detects new anomalies in real time, which can be used as a supplement to the existing detection rules.
- For analysis: Autonomous root cause analysis to reduce human effort.

7.2.3 Application and Performance

Application Aware RAN Optimization

ETSI standardization of Multi-access edge computing aims for a smooth interoperability between intelligent management platform and applications. It enables a plug and play architecture, helping the application developers, to create applications in an intelligence management platform vendor independent manner. This solution supports ETSI defined application aware APIs. Application aware APIs provide an interface and a framework for the edge computing applications to optimize RAN based on the Application request and report API's. Example: Based on QoE request and Report from gaming application, the intelligent management platform will do RAN optimization say by denying ENDC admission to low priority users, when the QoE of the high priority users are degrading.

- User experience Without RAN optimization

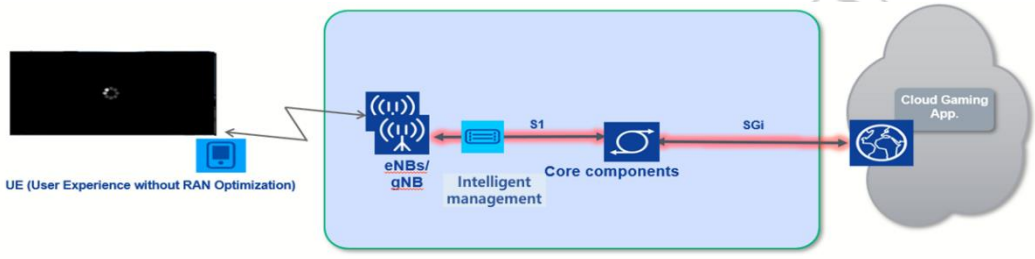


Figure 7-1 User experience became worse when cell load without RAN optimization

In this case when the Cell is loaded in the RAN or when the UE moves to cell edge, then latency and Jitter of the gaming user will increase to an extent which makes it impossible for the user to continue the gaming further and results in a very bad user experience.

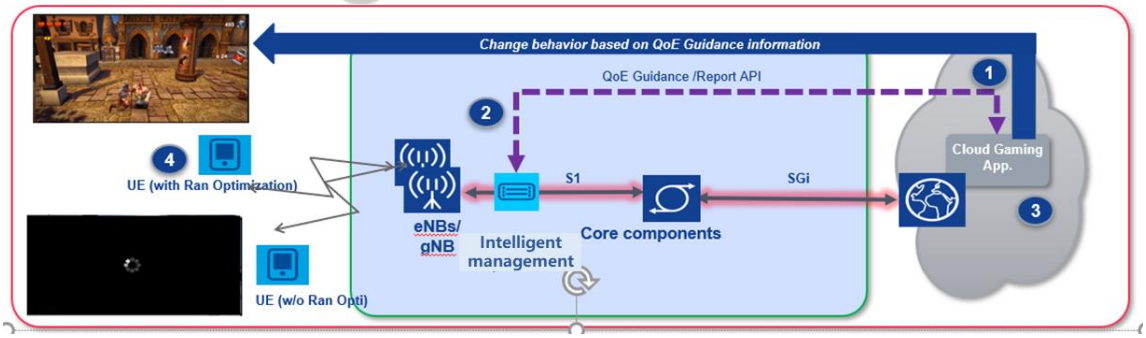


Figure 7-2 User experience is better with RAN optimization

- As part of this Application aware API interface, applications will be able to provide its QoE requirement towards intelligent management platform via QoE request API, where it will say its minimum and maximum requirement with respect to their key metrics. For example, in the gaming application some of the key metrics used are throughput, latency, jitter etc.
- Additionally, on a regular interval applications will also send their current QoE report of the

user session indicating the current live metrics like throughput, latency, jitter, etc.

- These API's ensure the needed information to be made available in intelligent management platform for the RAN Optimization.

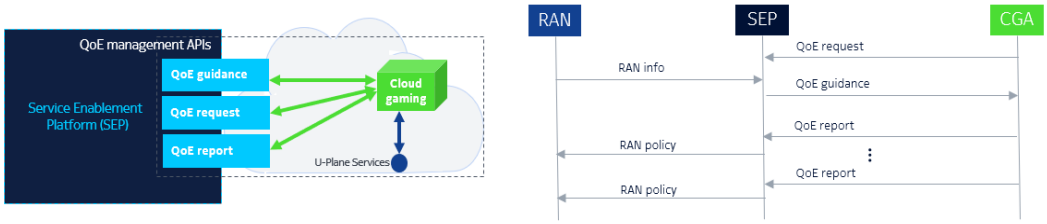


Figure 7-3 QoE optimization via Application aware API service

Based on this QoE request and QoE report values made available via Application aware API's in intelligent management platform from CGA, enables to build intelligence in intelligent management platform to decide and block certain user group, so that say the gaming users QoE will not go below the minimum QoE requirement provided during QoE request, hence the user will be able to play the game in each and every situation irrespective of the cell load, radio congestion, etc.

7.2.4 Summary

Network perception is the basic prerequisite for closed loop network services and automatic driving. The Network Autonomous perception including flowing 3 function blocks: Perception, Comprehension and Projection. According to the objects, events, environments of the user cases, Network perception implementation/solution can be classified as application perception, location perception, Network problem perception etc. The example use case of "Application Aware RAN Optimization" show benefits of making the application use cases more flexible with the Application Aware API's in the MEC/SEP.

7.3 Autonomous Diagnosis

7.3.1 Description

The core and difficulty of alarm processing or fault processing of communication networks is the diagnosis and location of fault causes. With the continuous development of communication technology, the scale of communication network is expanding, the type of network elements is increasing, the degree of network heterogeneity is improving, and the professional network management is gradually integrating, the difficulty and complexity of alarm analysis and processing are increasing under the condition of the continuous enrichment of alarm information. In the traditional alarm processing process, network maintenance professionals usually analyze, judge and locate the cause of fault based on the alarm information, and then determine how to deal with the fault, while autonomous fault diagnosis is based on artificial intelligence technology to confirm the alarm information through appropriate filtering, screening, matching and classification processes, and trace the source of alarms based on the relationship between the alarms, shield

unimportant or derived alarms, and achieve the goal of "alarm tracing". The network fault diagnosis can be realized quickly by filtering, screening, matching, classifying and other processes to confirm the alarm information, trace the alarms according to the relationship between the alarms, block the unimportant or derived alarms.

7.3.2 Implementation

Generating diagnostic rule base based on AI learning

- Diagnostic information acquisition

The richer the diagnostic information, the better the diagnosis effect, so firstly, it should have the function of automatically acquiring each diagnostic information of the whole cycle (current and historical) of the network. For example, in the operation of the existing network, in addition to collecting operation logs, alarms, KPI, fault handling suggestions and other daily monitoring data, the network topology, service configuration and service status, which only record the current state, should also be sampled regularly and used as the material for machine learning.

- autonomous diagnosis by AI algorithm

  - Extraction of fault features is to construct a lower dimensional feature space using existing fault feature parameters, map the useful information contained in the original fault features to a few features, and ignore other redundant and irrelevant information. Here, algorithms such as data dimensionality reduction and classification can be used. For example, the alarm data of the existing network is pre-processed, including data import and cleaning, redundant information removal, data degradation processing, frequent alarm identification, etc.

  - Analyze the causes and explore the root causes based on the alarm information, service configuration/status data, operation logs, fault handling records, and other related data during the period between fault generation and clearing. Correlation algorithm and deep learning algorithm can be used here. Typical association algorithms include Aprior algorithm, FP-G (Frequent pattern Growth) algorithm, FreeSpan algorithm and prefix span algorithm.

  - By analyzing enough cases, get all possible causes and calculate the probability of causes. Here the relevant algorithms of probability theory can be used.

The operation of diagnostic rules

  - Existing network monitoring: real-time monitoring of alarms, and regular sampling of traffic and packet loss, and recording of operation logs.

  - Matching fault characteristics for fault diagnosis: Matching the existing network monitoring data in real time, and once the matching is successful, the diagnosis starts immediately. The causes of faults are sorted by probability from largest to smallest and diagnosed one by one. When the existence of a cause is confirmed, the fault can be located and treatment suggestions are given. ,

- Fault repair confirmation, reverse correction of the diagnosis rule base: after the fault is repaired by automatic recovery or dispatch order, feedback whether the cause is valid in the dispatch order and correct the probability of the cause in the diagnosis rule base.

7.3.3 Application and Performance

Application Scenarios:

There are various business application scenarios for network fault diagnosis: fault diagnosis within the same network element, fault diagnosis of lower layer services on the same professional network, fault diagnosis of cross-professional network, integrated fault diagnosis, autonomous fault prediction, etc.

autonomous fault diagnosis process：

Based on the combination of big data association rule analysis and artificial intelligence technology, the autonomous fault diagnosis system solution outputs a series of rules between fault characteristics and fault causes based on the upstream and downstream relationships of the network and services in the system, synthesizing all monitoring data (including alarms and performance), operation logs and fault resolution history records.

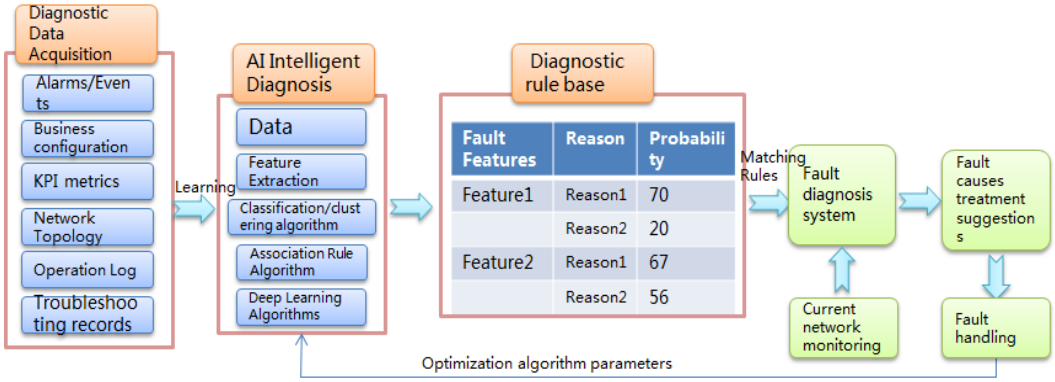


Figure 7-4 Diagnostic Intelligence Flowchart

7.4 Autonomous Prediction

7.4.1 Description of Predictive Intelligence

A prediction capability refers to the ability to perform an *estimate* of an unknown, often future, state based on some known dataset. As such it is a vital component of the autonomous network and its ability to respond to current and expected changes in the environment. For operators, vendors, and third parties the ability to do accurate, autonomous, automated predictions is crucial to maximize the performance and efficiency of the network while making best use of investments and minimizing costs.

Prediction and forecasting techniques have been used before the advent of Artificial Intelligence

and Machine Learning. Using the latter techniques provides benefits like more dynamic prediction on changing data due to AI/ML capacity to learn complex data, reducing the need for detailed and manual investigation by experts on every single prediction, which can reach into the millions. AI techniques are well suited for the complex, high-dimensional, and large data sets that signify the Telecommunications networks.

Prediction capability is a crucial component of the autonomous network and has strong interdependence with other intelligence capabilities. The perception capability is required to acquire a dataset on which predictions can be made and preventive actions are made possible through diagnosis based on predictions, supporting the ability to make autonomous decisions. The relationship between prediction capability and other capabilities is further described in chapter 7.6.

The prediction capability is defined here as the ability to forecast time series, long- and short-term, or to predict a best action, or best input to an action, based on current environment. These abilities are necessary at all levels of the network. From UE level predicting a best action, such as link-adaptation for optimizing spectrum utilization, to an area-wide prediction of network demand enabling cell sleep mode for best energy efficiency. Finally, long-term capacity planning can utilize long time-series prediction to maximize use of already installed base and the return on future investment in network capabilities. A pressing concern for operators is to reduce customer churn by managing the customer experience. In this case predictive capabilities play an important role to determine likelihood of customer change 错误!未找到引用源。.

From UE level to network-wide, and from milliseconds to days or even years prediction capabilities can significantly enhance the network ability and other autonomous capabilities.

7.4.2 Key Considerations for Implementation of Predictive Intelligence

Many modern tools for prediction use deep neural networks to take large scale datasets and, depending on method, use machine learning or machine reasoning with numerical or symbolic representation respectively to reach a conclusion. Abstraction of the data is done either implicitly by mapping to numerical values or more directly in machine reasoning.

There are multiple considerations when designing and using a prediction algorithm. Two important ones regarding the output are accuracy and granularity:

Accuracy is one way to determine how good a prediction is. Other methods include precision and recall which works well for classification problems. There are multiple methods for determining accuracy, like mean-squared error. This and others are often used in training to determine the loss function and thereby how to update the weights of the network. When executing the algorithm and to understand how useful it is another evaluation function can be used: prediction interval. A prediction interval is used to estimate the likelihood that a prediction will be within a certain distance of the true value. It is based on the estimate that the error, or distance between the prediction and the truth should follow a normal distribution. For a given prediction using a 95%

prediction interval there is a 95% chance that the distance between prediction and truth should be less than X. This is useful when doing predictions that may have impact on the network performance.

Granularity is a concept to define how fine grained the data and predictions are. In the case of time series predictions granularity can be described by time interval. A time interval is the shortest period of time in which datapoints are aggregated and one prediction can be made. Setting the time interval correctly is important and especially so when the prediction supports a decision or action, especially if there are latency requirements.

Since predictions rely on data the quality and utilization of that data directly relates to the quality of the prediction. With machine learning the training depends on data and so if the data used for training is poor or poorly reflects the data used in the inference/execution phase the resulting model will also perform poorly.

There are multiple questions that must be considered that relates to, or are impacted by, the data at hand. The size and numbers of variables in the dataset is a basic example, in the autonomous network there is a strong dependency here to the perception capability, which is further explored in chapter 7.2. For time series there are multiple specific data considerations that must be handled. One reason that the size or length of the data set is important is because the data may be subject to tendency and seasonality. The dataset must be sufficiently large so that the model is trained on these characteristics. The time interval is related to this issue, since a larger time interval may smooth out seasonality characteristics. An example would be predicting network usage in an area on a daily basis may not adequately prepare the network for peak rates that may occur during specific times of the day.

For the input data the time interval can often be chosen when deciding how to aggregate the datapoints, for example on second, microsecond, or daily basis. This also has an impact on the time interval of output, in best case they are matched. In a worst-case scenario, the step-size is larger than the granularity. Trying to for example to predict 1 second level output based on daily level data is going to be difficult at best.

Once the time interval is determined the input and corresponding output can be decided. This means deciding the number of time-steps used as input to produce the sought predictions. As must how many steps ahead those predictions are.

A final consideration is the difference between correlation and causation. When using machine learning differentiating between correlation and causation is not immediately obvious. This means that the model may induce relationships between datasets that are not reflected in reality. Creating a good algorithm for prediction therefore still requires involvement from subject experts.

7.4.3 Application and Performance

7.4.3.1 Capacity Prediction for Network Planning and Evolution

Capacity Planner is a solution that gives CSP the unique ability to better understand the CAPEX and OPEX impact of introducing new services, and the general network dimensioning process.

To achieve this goal, Capacity Planner is based on a combination of Traffic Forecasting with Utilization KPI prediction based on state-of-the-art Machine Learning techniques. This enables the detection of capacity problems in a proactive way, leading to optimum bottleneck identification in a pre-defined time frame as opposed to the traditional way of working based on the reactive mode when Utilization/Capacity KPI exceed a pre-defined threshold.

Capacity Planning is an AI-powered solution that helps the radio engineers in the process of obtaining optimized network dimensioning to achieve most efficient use of the capacity, for current and future traffic. Capacity Planning support LTE/NR network technologies. The AI-enabled prediction engine includes a traffic forecasting module that provides at cell level/sector level traffic category KPI forecast, allowing dimensioning for future traffic. Capacity Planning also provides optimal dimensioning actions to avoid traffic bottleneck. These actions include traffic balancing, addition of new carriers or sectors, and estimation of new site additions. The time interval for capacity planner could be hourly, daily, or weekly. Nowadays monthly level prediction is rarely used and mainly depends on the long-term monthly level of input raw data.

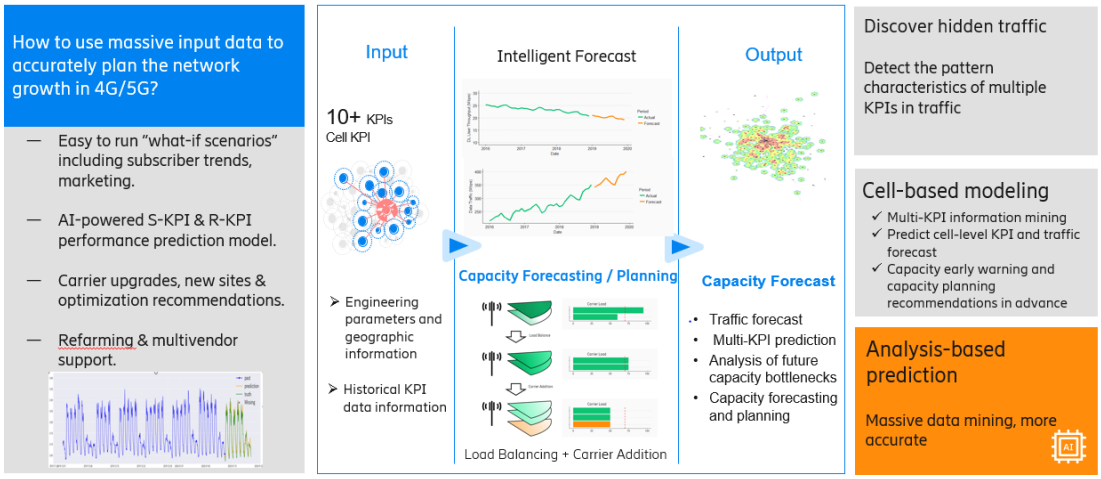


Figure 7-5 Function and Application

7.4.3.2 BLER Target or MCS Prediction for Link Adaptation

LTE/NR DL spectral efficiency may be affected by inter-cell interference. The current DL Link Adaptation takes inter-cell interference into account, but it is optimized towards slowly varying interference scenarios. This might result in suboptimal performance in scenarios with burst traffic in the neighboring cells when interference impact comes in highly varying short bursts.

To increase spectral efficiency by improving how Downlink Link Adaptation adapts to inter-cell interference scenarios created by burst traffic, additional Machine Learning algorithm can be introduced for steering of the existing Link Adaptation. The ML algorithm is trained to recognize burst interference scenarios based on the neighbor cell activity and UE signal quality.

BLER target with the highest spectrum efficiency for each UE can be predicted using ML model and the predicted BLER target can be set for each UE dynamically during link adaption. The time interval for BLER target prediction and setting is in the level of several hundred milliseconds.

Another choice to use ML in link adaptation is to use ML model to predict MCS value directly for

each UE, the time interval for MCS prediction and setting is in the level of milliseconds.

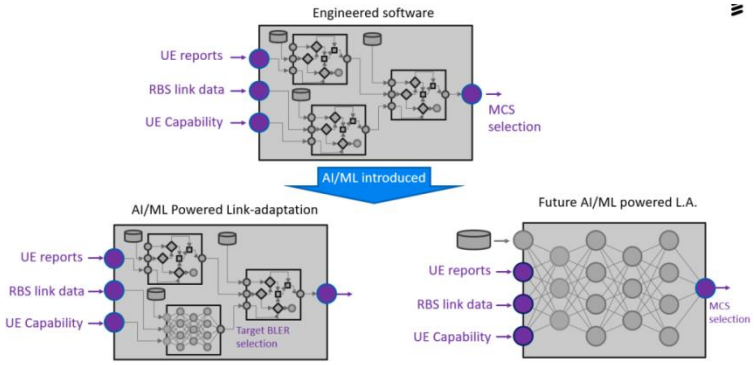


Figure 7-6 AI/ML powered link-adaptation

7.4.3.3 UE Coverage Prediction for Mobility Control

It is a trend that more and more operators deploy multiple frequencies in their network. Usually, not all frequencies have the same coverage on all places in the network and there are those operators that would like to configure each base station the same, independent of if there is coverage from other frequencies or base station. Today's implementation of mobility control functionality(similarly for other functionalities as load balancing and s-cell selection) is not optimal in the sense that they use information on what is configured but could not actually present for a specific UE location, so:

- Machine learning can come into place: learning the correlation between a UE's position and existence of other carriers enables the prediction of UE coverage to support mobility control functionality, making better selection on frequencies for a UE.
- UE's position here is not necessarily as a geographical position, it is a rather relative position within the cell. See 错误!未找到引用源。, where it's important to not learn the position of each area but rather the radio properties of each to distinguish them;
- UEs located in the area could be moved to frequency F3 or F4 but would not move to F2, so the measurement for F2 can be unnecessary based on model prediction results.

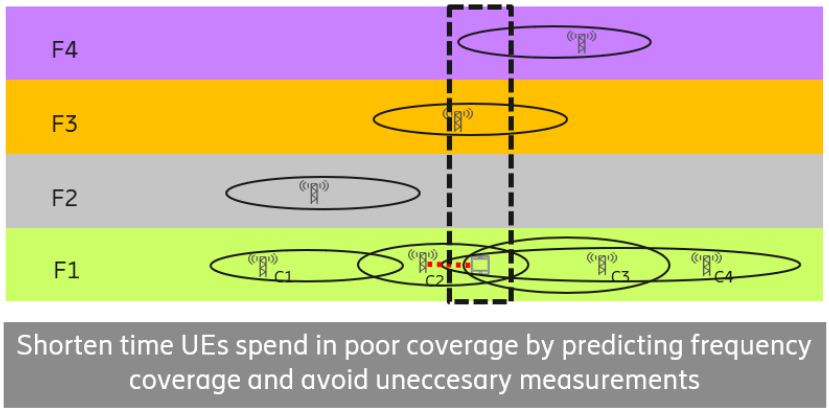


Figure 7-7 UE coverage prediction for mobility control

7.4.3.4 AI Powered MIMO Sleep Mode

Prediction is the most important capability in enabling MIMO sleep mode. To get more power saving, machine learning is a technique that finds patterns in data, and then given those patterns, **makes predictions**. Good and suitable predication on traffic will bring more power saving on this function

To train a machine learning model such that an algorithm can use it to make predictions, we need labeled data. Labeled data is a dataset such that every data item contains feature values and a corresponding prediction. We usually start with raw data which are collected from real network and perform several data processing steps to transform it into a labeled data set. Based on the labeled traffic data with a long period, we can train a model with algorithm to make predictions of future traffic.

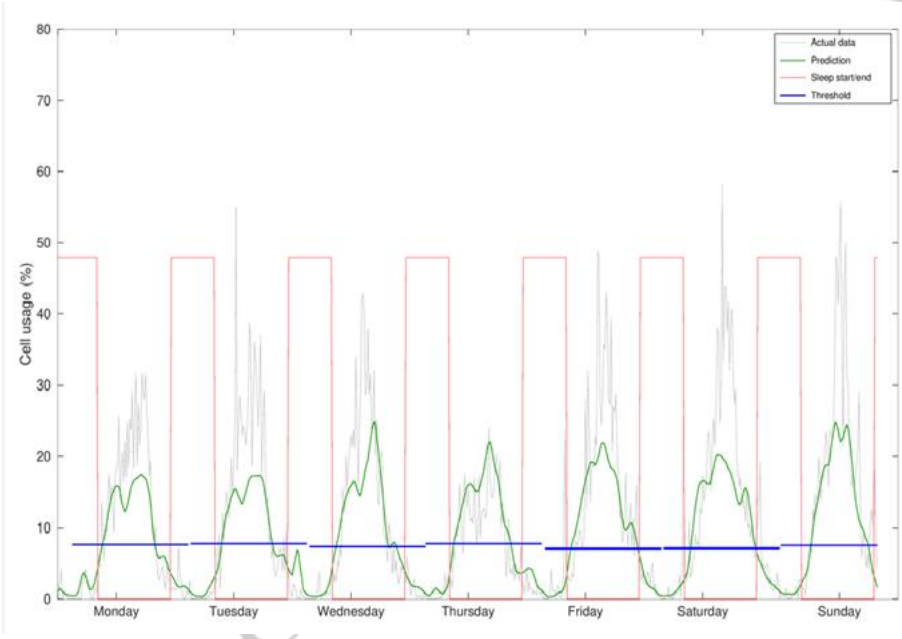


Figure 7-8 AI powered MIMO sleep prediction

7.5 Autonomous Control

7.5.1 Overview

Autonomous control of wireless networks is implemented at the execution layer of autonomous networks. It can be classified into the following functions:

- Network-level autonomous control:

Based on the analysis of network data and diverse scenarios such as user distribution and cell load, the EMS uses artificial intelligence algorithms such as reinforcement learning to sense the surrounding environment for the wireless network and provide the most suitable parameter combination. In this way, the EMS dynamically adjusts and controls the network through functions

including load balancing parameter optimization, handover parameter optimization, cell-level scheduling, power control parameter adjustment, channel-level optimization, and dynamic cell or carrier shutdown.

- Grid-level autonomous control:

In the future, site will be capable of high real-time resource autonomy based on intent. Based on the intention information transmitted by the upper-layer EMS, real-time wireless network resources, and surrounding environment factors, the base station performs real-time and accurate scheduling of radio resources at each layer by using the native intelligent prediction capability, maximizing network performance and user experience. For the intent control information, it includes requirement for service, energy saving and so on. Real-time resource status and surrounding environment factors includes coverage, spectrum efficiency, energy efficiency, load, interference, and terminal capability. And resource scheduling at each radio layer is referring to spectrum, power, and channel.

7.5.2 Application and Performance

7.5.2.1 Use Cases of Network-level Autonomous Control

Virtual grids represent signal strength, as shown in the following figure. Unlike traditional geographic grids, virtual grids are not classified geographically, but represent the reference signal received power (RSRP) of multiple cells. They are used for carrier selection, carrier signal estimation, and spectral efficiency estimation in inter-frequency scenarios.

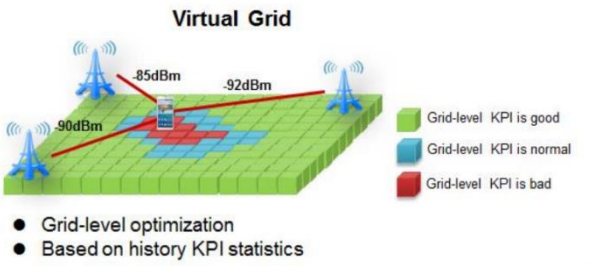


Figure 7-9 Virtual grid

There are seven multi-band parameters and handover parameters, each with about 10 values. As a result, optimization space is large. Traditionally, virtual grids are constructed in tables, with each row indicating one virtual grid. If the collected samples are not sufficient, the base station may incorrectly determine the target frequency for handovers.

Application Scenarios

In multi-band scenarios, cell reselection parameters need to be properly set to balance the number of UEs between cells operating on different frequencies. In most cases, high-band cells (L2600) are recommended for cell center users (CCUs), and low-band cells (L1800) are recommended for cell edge users (CEUs). The number of UEs camping on a cell should be determined with the aim of helping to maximize the sector-level UE throughput.

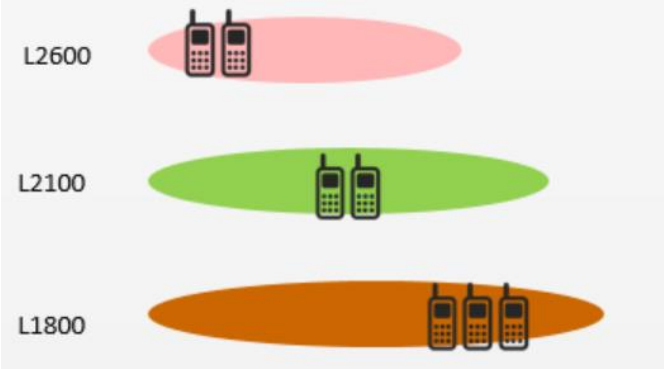


Figure 7-10 Multi-band cell reselection

Different reselection parameters are set for these cells so that UEs in idle mode can camp on different cells, achieving load balancing between these cells.

Solution Overview

This solution enables the EMS to automatically optimize the combination of cell reselection parameters and handover parameters in multi-band scenarios. Based on UE-level MR data, the EMS traverses the value ranges of the parameters, and then builds a data model based on historical data to estimate the frequency bands that these UEs can camp on as well as the UE throughput. In this way, the optimal parameter value combination is obtained. The following figure shows the estimation process.

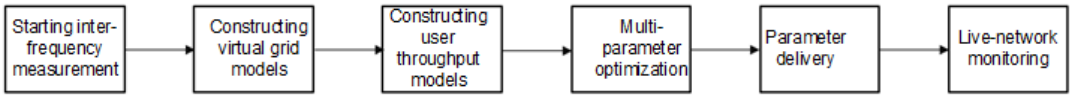


Figure 7-11 Process of the multi-parameter optimization solution

Based on the combination of handover and idle-state parameters for cells on different frequency bands, the base station constructs a virtual grid model to predict inter-band UE transfer and changes in SINR, as well as UE throughput. Based on the number of UEs and bandwidth in the cell, the EMS calculates the multi-band parameter combination that delivers the optimal overall spectral efficiency and UE throughput.

Based on inter-frequency measurements, the base station constructs a virtual grid model which it uses to predict the UE throughput after handovers. The base station can obtain the signal strength, traffic, and throughput of UEs on frequency band F1 based on MR data. After the reselection or once handover parameters have been modified, if the UE is reselected to frequency band F2 based on the reselection rules but the signal strength of F2 is unknown, the base station can calculate and predict the throughput of the UE on F2 based on the virtual grid model and the UE throughput model. After the parameters are issued, UEs can be automatically handed over to the cell with the optimal throughput experience, improving both the handover success rate and UE throughput.

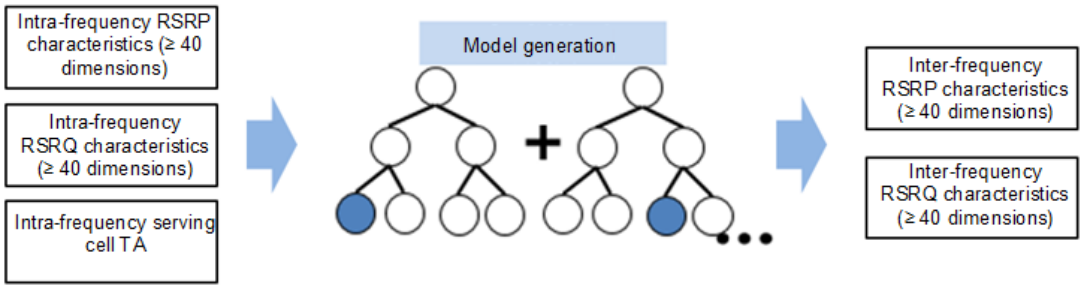


Figure 7-12 Virtual grid model building

Effect

This method has been fully demonstrated at an actual site. The following figure shows UE throughput changes after optimization, with the red line indicating optimized data. You can see that the UE throughput increases significantly after the optimization.

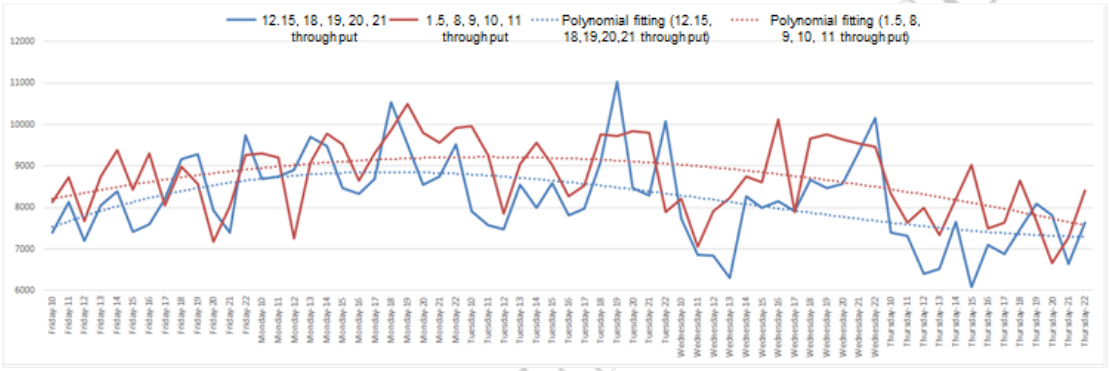


Figure 7-13 Multi-parameter optimization verification result

7.6 Relationships among the Essential Capacities

The interplay and dependency between the different capabilities can be envisioned as the diagram below:

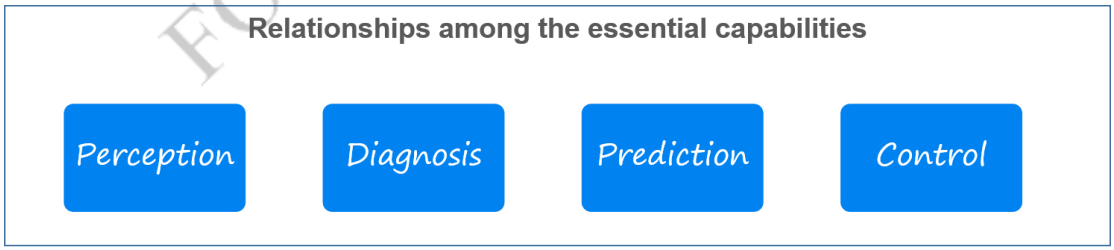


Figure 7-14 Relationships among the essential capabilities

The four essential capacities are independent to each others but will be effected mutually in the algorithm, data, computation power and other O&M domains.

Perception for instance can feed directly into any of the other capabilities. Some examples of alternative linkages between the capabilities as compared to the diagram are listed here (non-

exhaustive). There is significant interdependence between prediction and diagnosis as a diagnosis can lead to a new prediction, and of course there is a larger loop in play as exemplified by the link adaptation in which case a prediction leads to a decision which then affects further predictions. Further, due to the large amount of data that a network generates filtering will be required regarding what data to collect. In the future it is possible that decisions can be taken autonomously by the network regarding what data to collect, thereby impacting the perception capability. Perception should in turn be able to directly feed data to every autonomous capability. Predictions should be used directly by decision and control capabilities for proactive decisions and actions.

7.7 Summary

The autonomous capacities have been refined the autonomous O&M scenarios into four types, that is, autonomous perception, autonomous diagnosis, autonomous prediction and autonomous control.

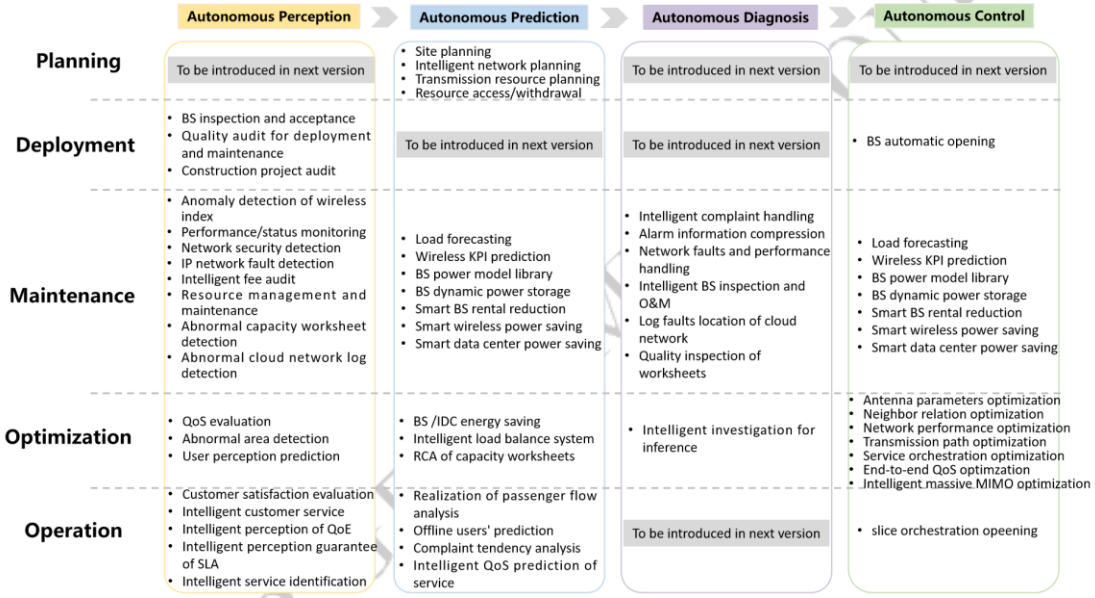


Figure 7-15 Use cases illustration between essential capacities and life-cycle dimension

The use cases within both essential capacities dimension and life cycle dimension are illustrated above, with the four essential capacities of autonomous O&M, it is clear that the innovation pilot and achievement promotion of AI technology in image recognition, big data analysis, complex computing and other O&M domains, AI technologies have been gradually applied in all the processed of O&M life cycle.

8 Autonomous Network Level

8.1 Introduction

Different autonomous mechanisms in the telecommunication system may lead to different capabilities of intelligence and different operation efficiency on network management and control workflow, and indicates the level of the autonomous network. Autonomous network level describes the level of application of intelligence and automation capabilities in the network management and control workflow. Participation of the human and telecommunication system in the network management and control workflow are different for each level and are important factors to evaluate the autonomous network level. For each autonomous 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 autonomous network level, all tasks are performed by telecom system.

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

8.2 Framework Approach for Classification of Autonomous Network Levels

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

- “Human” represents corresponding tasks are accomplished by human or human utilizing the tools for network and service management and orchestration.
- “Human & System” represents corresponding tasks are accomplished by collaboration of human and telecom system, the detailed collaboration pattern depends on the scenario, which is not addressed in the framework approach for evaluating autonomous network levels.
- “System” represents corresponding tasks are fully accomplished by telecom system.

Note: the following framework approach for classification of autonomous network level is based on the framework approach for classification of autonomous network level defined in 3GPP TS 28.100. And the framework approach for classification of mobile network management and operation intelligence levels defined CCSA TC7.

Table 8-1 Framework approach for classification of intelligence network level

Intelligence Network level		Task categories				
		Execution	**perception**	**Analysis**	**Decision**	**Intent handling**
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary autonomous network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate autonomous network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system
Note 1: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision. *Note 2: The order of above five task categories does not reflect the workflow sequence.*						

- **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 perception 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 perception efficiency.
- **Level 2 preliminary autonomous network:** All the execution tasks are accomplished automatically by telecom system itself. A part of the perception 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 autonomous network:** All the execution and perception 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 autonomous network:** All the execution, perception, analysis and decision tasks are accomplished automatically by telecom system itself. And intent handling tasks can be partly accomplished automatically by telecom system itself based on human defined intent handling policies. At this level, telecom system can achieve the intent driven closed loop automation based on human defined intent handling 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 handling policies.

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

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

For the classification of autonomous network, we have done some research in White Paper 1.0, and output some research results to give the current level of intelligent classification for some typical cases. According to the analysis, most of the current use cases in the industry are between level 2 and level 3. For example, in order to evolve to level 4 in the case of coverage optimization, 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.

For another example, intelligent energy efficiency enables closed-loop overlay optimization automation driven by Level 3 energy efficiency policies. In order to achieve the evolution to a Level 4 autonomous network, energy saving monitoring rules, optimization requirements and energy saving policy determination, network/service assurance intent assessment and performance deterioration prediction all need to be automated. Another example is the NR UE throughput optimization, which automates the closed-loop NR network UE throughput optimization driven by Level 3 policies. To enable the evolution to a Level 4 autonomous network, the RAN manager also needs to predict potential UE throughput problems and automatically generate UE throughput optimization policies based on wireless network/service assurance intent. At Level 4, intent-driven, closed-loop coverage optimization automation based on scenario-specific service assurance intent will be implemented.

Subsequent chapters of this white paper will continue the study of autonomous tiering for selected use cases, giving the evaluation of tiering and the goals of subsequent tiering evolution.

8.3 ANL Evaluation of Typical Use Cases

8.3.1 ANL Evaluation in Commercial Network

Based on the field network workflow for network optimization, the execution, awareness, analysis, decision, and intent handling are further divided into specific tasks in the workflow, and the grading

criteria of each task are defined. The two-step evaluation method is used for ANL evaluation. The inter-generational evaluation method is used for preliminary evaluation, and detailed evaluation is performed then based on weighted evaluation to obtain the intelligence level score of the case.

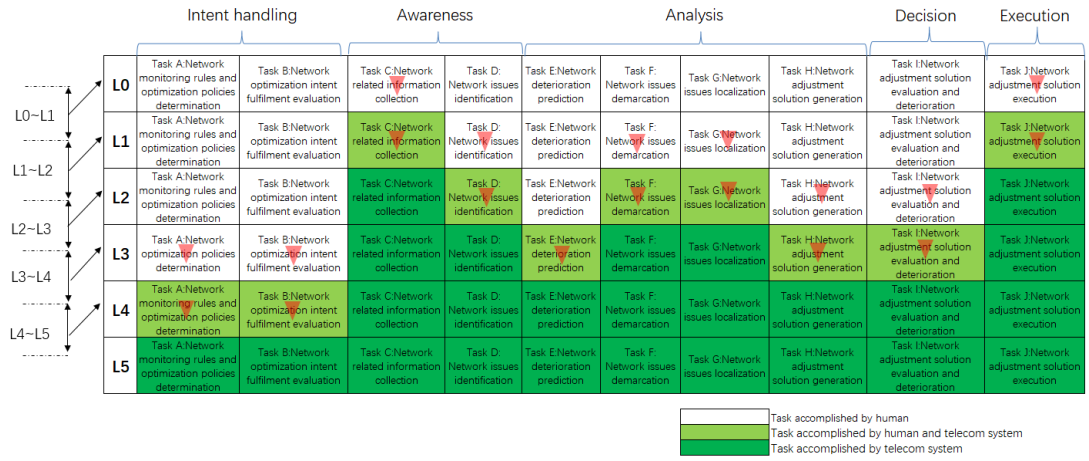
8.3.1.1 Evaluation Method

Step 1: As shown in the following figure, white indicates that the task is manually performed, light green indicates that the task is performed by both humans and the system, and green indicates that all operations are performed by the system. Determine the automation status of each task, and then obtain the case level or level range based on the following figure. If the inter-generational evaluation result is an integer, weighted scoring is no longer required for the case. For example, if the inter-generational feature evaluation result is L2, task scoring is no longer required and the final result is L2.

Step 2: If the inter-generational evaluation result is within a range, for example, between La and Lb, the tasks for which La and Lb criteria are different are scored. The following figure shows the tasks that differ among levels. Assume that the tasks for which La and Lb criteria are different include task 1, task 2, and task 3. These tasks are determined in sequence. The score of a task that meets the criteria of Lb is 1, and that of a task that does not meet the criteria of Lb is 0. The scores of the three tasks are obtained: s1, s2, and s3.

Step 3: Assume that the weights of tasks 1 to 3 are w1, w2, and w3, respectively. The weighted average method is used to obtain the score of each case, that is, La + (s1 x w1 + s2 x w2 + s3 x w3)/(1 x w1 + 1 x w2 + 1 x w3).

Table 8-2 Level evaluation of some commercial province network



8.3.1.2 Task Weight Setting

The task weight is set based on the manual time required and technical complexity. In addition, the time weight must not exceed the highest weight of the technical difficulty of the task (use case).

Assume that the technical difficulty is divided into N levels in ascending order, each level is numbered n, there are M tasks, and the time consumption proportion of each task is T_m%. Set the benchmark weight (that is, the weight is 1) for the task that takes the least time and has the lowest

technical difficulty. When the technical difficulty is the lowest, the task weight is T_m/T_1. Each time the technical difficulty is increased by a level, the overall weight is increased by 1 on the basis of the time weight. The weight can be set based on the following table.

Table 8-3 The weight of task time consumption

Time / Difficulty	T_1%(No. 1)	T_2%(No. 2)	T_3%(No. 3)	T_4%(No. 4)	T_M%(No. *M*)
1	1	T_2/T_1	T_3/T_1	T_4/T_1	T_M/T_1
2	2	$T_2/T_1 + 1$	$T_3/T_1 + 1$	$T_4/T_1 + 1$	$T_M/T_1 + 1$
3	3	$T_2/T_1 + 2$	$T_3/T_1 + 2$	$T_4/T_1 + 2$	$T_M/T_1 + 2$
...	...	...	...	...	...
N	N	$T_2/T_1 + N - 1$	$T_3/T_1 + N - 1$	$T_4/T_1 + N - 1$	$T_M/T_1 + N - 1$

Note: You are advised to round off the division result to an integer.

In addition, in consideration of the restriction that the time weight round(T_m/T_1) must be less than or equal to N, the final weight is summarized into the following formula:

$$\text{Weight} = \min[\text{round}(T_m/T_1), N] + n - 1$$

8.3.1.3 NR Coverage Optimization Evaluation

Wireless network coverage optimization uses configuration management (CM) and performance measurement (PM) data on the network to identify problems on the live network, analyzes root causes, and provides optimization solutions. The following table describes the basic capability of each task of wireless network coverage optimization. The inter-generational evaluation result shows that the case level is L2 to L3. Then, level evaluation is performed on the tasks for which the L2 and L3 criteria are different to obtain the scores of these tasks. In addition, the technical difficulty is classified into four levels in ascending order. The time proportion of each task is obtained based on the actual time consumption. The weight of each task is calculated based on the formula in section8.3.1.1. The following table lists the scores and weights of tasks for which the L2 and L3 criteria are different.

Table 8-4 The scores and weights of tasks for which the L2 and L3

Task	Current Solution Capability	L2 to L3 Scoring	Weight
Generation of coverage monitoring rules and optimization policies	Problems can be automatically identified and analyzed and optimization advice can be provided based on user-defined coverage and throughput targets and optional optimization methods (RF and parameter).	N/A	**5**
Wireless coverage assurance intent evaluation	The values of certain counters before and after the optimization can be compared.	N/A	**3**
Coverage data collection	The following functions are supported: online collection, automatic interconnection of engineering parameters, automatic generation of antenna files, and automatic MR collection.	N/A	**2**
Coverage performance exception identification	Cells with weak coverage, interference, and poor performance can be automatically identified based on preset target thresholds and parameters.	1	**2**
Coverage performance deterioration prediction	During optimization solution generation, deterioration analysis and prediction can be performed based on collected data.	N/A	**4**
Coverage problem demarcation	Currently, weak coverage, interference, and poor performance problems can be demarcated and analyzed.	1	**2**
Coverage problem locating	Currently, root causes of deep weak coverage, local weak coverage, overshoot coverage, and pilot pollution can be analyzed.	0	**5**
Coverage optimization solution generation	Currently, solutions to weak coverage, interference, and poor performance problems can be generated.	1	**7**
Evaluation and decision-making of the coverage optimization solution	The system automatically evaluates the performance of candidate solutions and provides solution selection advice. The optimization solution is determined manually.	1	**4**

Implementation of the coverage optimization solution	MML commands can be generated online and delivered in automatic, manual, or scheduled mode.	N/A	**1**

Calculate the weighted average value based on the preceding level evaluation result to obtain the level of the coverage optimization use case. The formula is as follows: S = 2 + (1 x 2 + 1 x 2 + 0 x 5 + 1 x 7 + 1 x 4)/(1 x 2 + 1 x 2 + 1 x 5 + 1 x 7 + 1 x 4) = 2.8. With this, the ANL of this case is evaluated as L2.8.

8.3.2 Wireless Broadband Service Provisioning

Based on the framework approach for classification of autonomous network level, the use case of Automatic Wireless Broadband Service Provisioning defined in clause 6.7.2 achieves the capabilities of level 3 autonomous network:

- Intent management (Human): performed by human.
- Perception (Telecom System): The RAN Manager automatically collects network operational data, such as RSRP, CQI .etc, and generate service provisioning map underpinned with AI based radio propagation modeling based on the collection results.
- Analysis (Telecom System): The RAN Manager can automatically analyze the collected data and generate provisioning map with 92% accuracy. With AI propaganda model and online simulation, service provisioning map could be automatically updated and refreshed based on latest state of the network.
- Decision (Human & Telecom System): By inputting the name of the target address, the pre-sale evaluation result can be given automatically for customer to make the decision, including whether or not the address is able to provide FWA service, what kinds of data package is available and what kinds of CPE are recommended.
- Execution (Human &Telecom System): If connected with the operator's inventory system and service system, the wireless broadband service operation system will start the logistics of the CPE and SIM card automatically once the customer places an order.

Table 8-5 Level evaluation of Automatic Wireless Broadband Service Provisioning

Intelligence Network level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent handling
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary autonomous network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate autonomous network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system

Note 1: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.

Note 2: The order of above five task categories does not reflect the workflow sequence.

To evolve to a level 4 autonomous network, the RAN Manager needs to provide the autonomous capability of Auto-balancing of multi-service, automatic value areas identification and network planning recommendation based on network problems forecasting.

8.3.3 Root Cause Analysis of Cell Performance Issue

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

- Intent management (Human): performed by human.
- Perception (Telecom System): After the automatic input of multi-dimensional feature information, it can then pass through the autonomous analysis engine module, without the need for manual experts to set rules, and can perceive and intelligently output the autonomous classification and identification of abnormal communities and problems.
- Analysis (Human & Telecom System): Combining the KPI and MR information on the wireless side, the autonomous 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 8-6 Level evaluation of Root cause analysis of cell issue

Intelligence Network level		Task categories				
		Execution	Awareness	Analysis	Decision	Intent handling
L0	Manual operating network	Human	Human	Human	Human	Human
L1	Assisted operating network	Human & Telecom system	Human & Telecom system	Human	Human	Human
L2	Preliminary autonomous network	Telecom system	Human & Telecom system	Human & Telecom system	Human	Human
L3	Intermediate autonomous network	Telecom system	Telecom system	Human & Telecom system	Human & Telecom system	Human
L4	Advanced autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Human & Telecom system
L5	Full autonomous network	Telecom system	Telecom system	Telecom system	Telecom system	Telecom system

Note 1: Human reviewed decision have the highest authority in each level if there is any confliction between human reviewed decision and telecom system generated decision.

Note 2: The order of above five task categories does not reflect the workflow sequence.

Subsequent evolution to level 4 autonomous network, the autonomous root cause analysis system for abnormal communities needs to realize the intelligence and automation of the further analysis part on the current basis. Several typical scenarios can be selected to test closed-loop function to realize automatic parameter optimization and execution, and finally lead cell performance improvement. In the long run, the autonomous RCA system can be combined with recommendation sub-system of the expert system or further integrated into the knowledge graph framework for full telecommunication system-level analysis and decision-making; and does not rely on manual intervention can automatically and autonomously give decisions, forming an optimized closed-loop automation.

9 Autonomous Network Architecture

9.1 Introduction

In 5G we expect AI/ML to be used at all levels, from an autonomous NE, to autonomous 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 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.

9.2 Framework Architecture of Autonomous Network

To gradually achieve the goal of a fully 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.

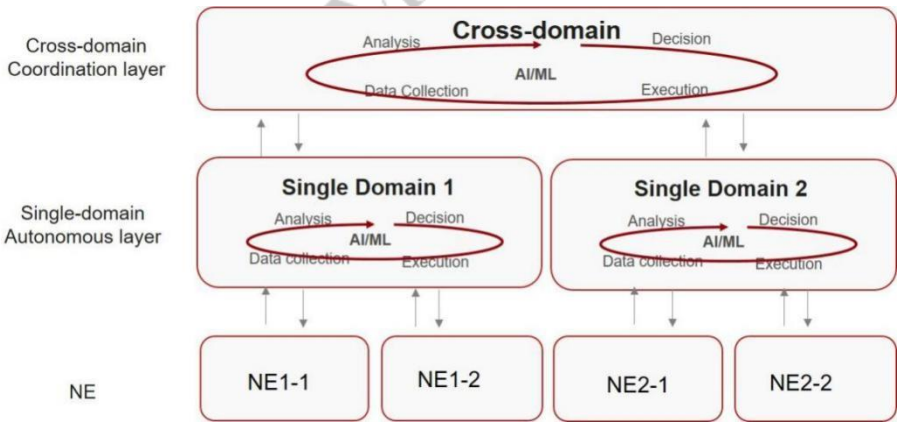


Figure 9-1 General network architecture for autonomous network

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.

9.2.1 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 9-2 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.

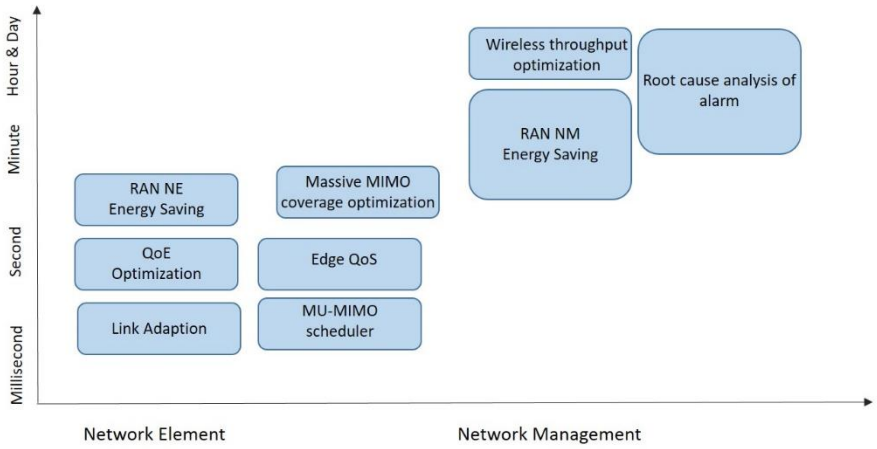


Figure 9-2 classification of exemplary use cases

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.

9.3 Architecture of Specific Use Cases

9.3.1 QoE Optimization

The example of QoE optimization architecture is shown in Figure 8-3. For the “QoE Optimization” use case, the autonomous management function is deployed to manage autonomous 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” performs 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.

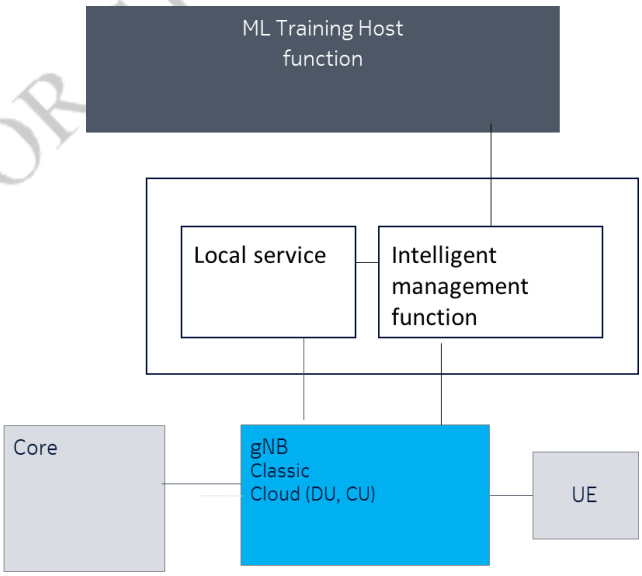


Figure 9-3 QoE optimization architecture (example)

9.3.2 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.

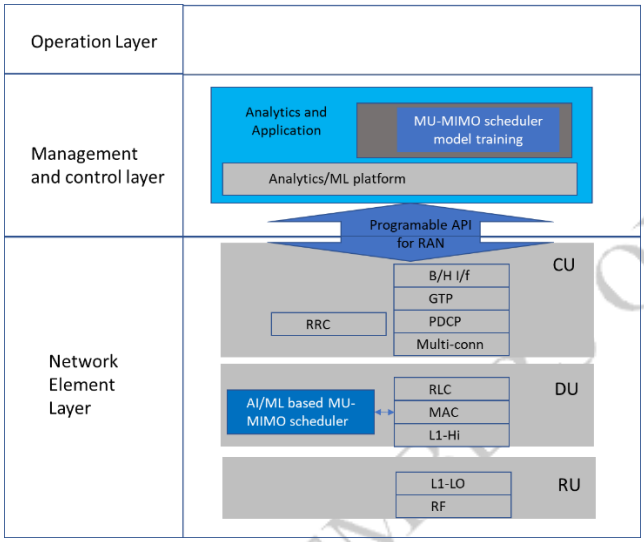


Figure 9-4 ML-based MU-MIMO scheduler architecture

9.3.3 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:

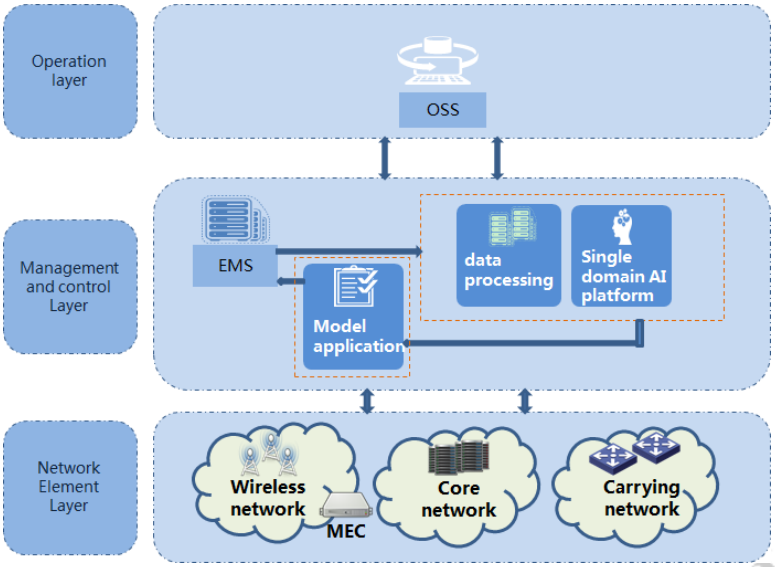


Figure 9-5 Root cause analysis of alarm Architecture

- **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.

9.3.4 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.

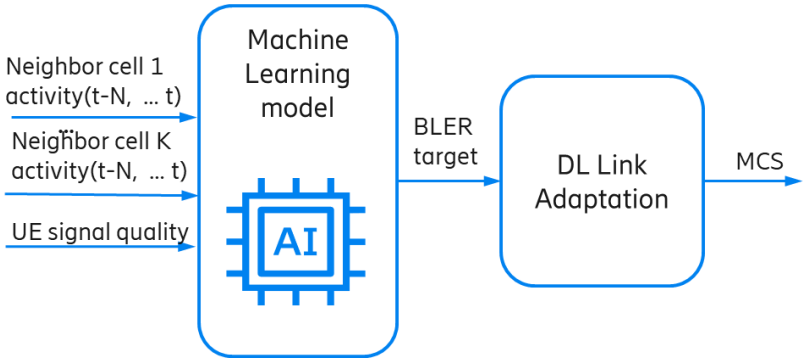


Figure 9-6 Process of AI based DL link adaptation

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 homogeneous 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.

9.3.5 Energy Saving

As shown in Figure 9-7, 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.

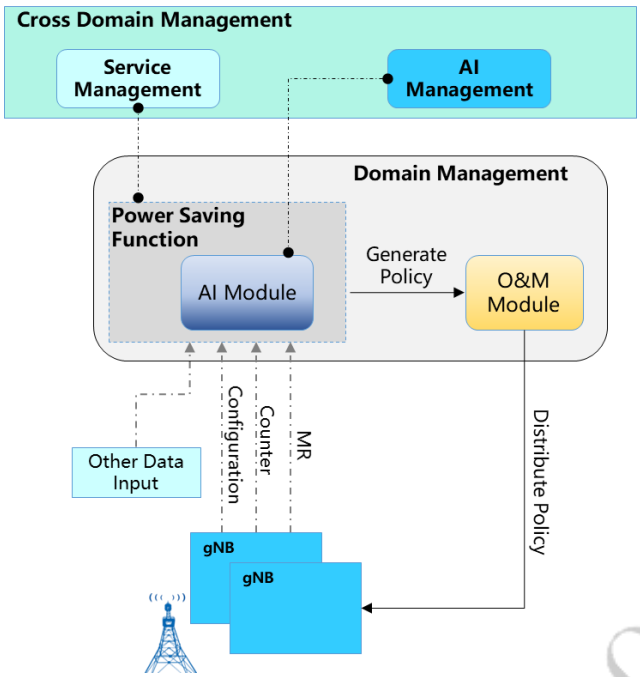


Figure 9-7 Three Layers Energy Saving with AI

9.4 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 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.

10 Autonomous Network Elements

10.1 Multi-Dimensional Data Collection and Reporting

As the main component of a network, Network Element needs to provide richer operation data with more complete structures, and provide an automatic collection and reporting mechanism to support the autonomous improvement of cases.

As wireless base station equipment, Network Element needs to provide richer and more complete data, and provide automatic data collection and reporting functions, to support the development of autonomous network.

10.2 Autonomous Data Modeling

Currently, Most of Network Element data comes from the traditional design, such as alarms, performance, and configuration.

With the increase of data volume and complexity, the weak association between data causes the difficulty of autonomous use cases in data cleansing and analysis.

Network Element needs to model data from the perspective of intelligence, to improve the relationships between different types of data.

10.3 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 autonomous use cases. With the enhancement of storage and computing hardware capabilities, it is also possible to store and calculate AI data on Network Element.

Network Element needs to provide small-scale data storage capabilities and AI computing capabilities, to support the implementation and deployment of autonomous use cases with real-time requirements.

AI model interaction and closed-loop control

Network Element needs to support interactions of AI model with the autonomous management function, including interactive interfaces such as AI model download, loading, rollback, and query.

Network Element needs to establish feedback and closed-loop control interfaces with the autonomous management function, such as parameter delivery for optimization and real-time calculation feedback, to implement autonomous case management and effect optimization.

11 Autonomous Network Management

11.1 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.

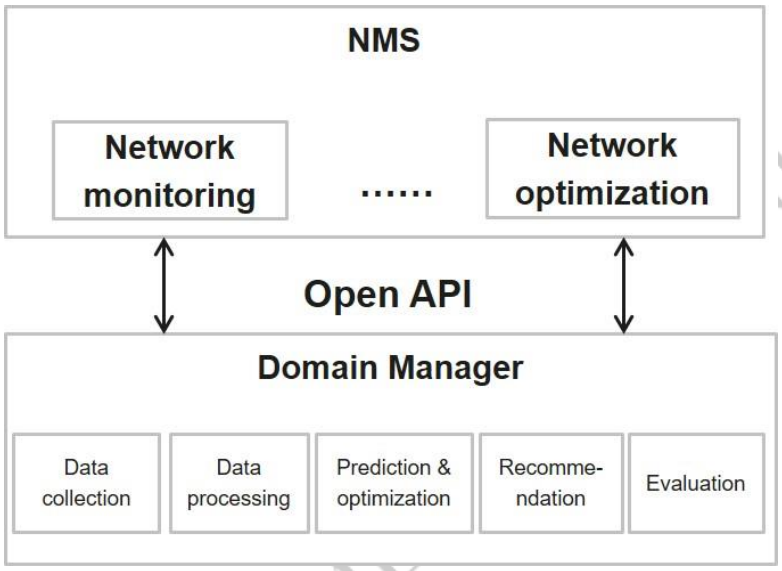


Figure 11-1 Network framework in a typical operator application scenario

11.2 Function Requirements for Network Management

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

11.2.1 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.

11.2.2 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.

11.2.3 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.

12 Further Studies

12.1 Trust in Autonomous Networks

12.1.1 Overview

12.1.1.1 Introduction

Mobile networks are evolving into the intelligence era with multiple application scenarios, features, services and operation requirements. Technologies including AI are expected to enable autonomous networks (AN) in areas such as network planning, deployment, operation, optimization, service deployment, and assurance. With the development of network systems and evolution of AI technology applications, self-X properties (the abilities to monitor, operate, recover, heal, protect, optimize, and reconfigure themselves) are expected to be achieved in AN, which means operators are supposed to gradually handover part of their works and duties to network systems themselves. AN becomes an inevitable trend of the network evolution. [7]

From the perspective of operators who are interested in and plan to deploy AN, the following questions may arise:

- Should an operator trust the AN?
- How to make human operator trust AN, and willing to hand over the control authority of the network to the AN system?
- How much can human operators trust their AN?
- What level of trust is sufficient/required for different mechanisms/aspects in AN?
- What are the most important factors for AN to earn human operator's trust?
- How to evaluate human operator's trust in AN?

If the above questions cannot be answered properly, the following serious problems and challenges will happen: the AN solutions including reliable ones and risky ones are difficult to be distinguished and may not be applied in operators' real networks due to the lack of confidence from operators, especially when the operators are facing the pressure of network quality assessment and market competition, traditional but more familiar solutions may be preferred then. On the other hand, from the AN's perspective, due to the lack of sufficient trust, AN solutions cannot get enough application opportunities in the real network, and thus loss the chance to learn and evolve to a more advanced level.

To solve the above problems, it is recommended to investigate and study trust in AN to help the telecom industry including communication service providers (CSPs),vendors and other industry participators to reach a consensus and unified understanding of the concepts and evaluation method on trust in AN.

12.1.1.2 Concepts

Trust is considered as a computational value depicted by a relationship between trustor and trustee, described in a specific context and measured by trust metrics and evaluated by a mechanism.

- **Trust**: the measurable belief and/or confidence which represents accumulated value from history and the expecting value for future. [65]
- **Trusted AN**: the autonomous network which is trustworthy enough (i.e. working correctly as intended), so that the network itself can be partly or completely autonomous.
- **Trust in AN**: a measurable and quantifiable degree of the trustor's confidence in a network to let it be governed by itself with minimal to no human intervention.
- **Trustor in AN**: the one who/which has the authority to authorize a network and/or the relevant entity be governed by itself with minimal to no human intervention.
- **Trustee in AN**: a network and/or the relevant entity with autonomy capabilities which is to be authorized to govern itself with minimal to no human intervention.

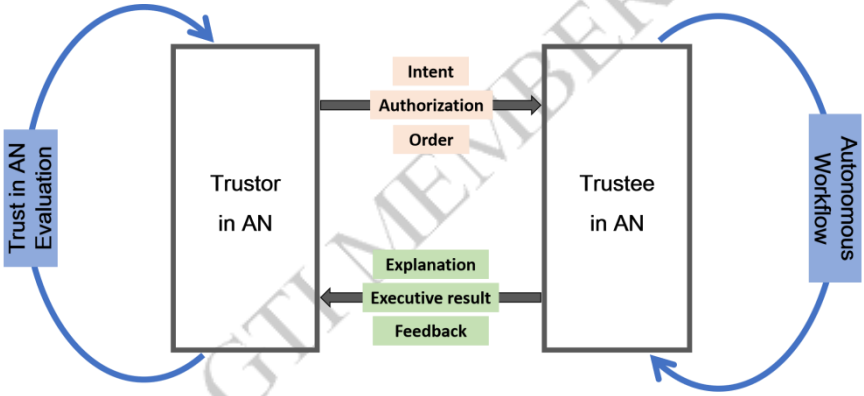


Figure 12-1 A conceptual model of trusted AN

12.1.1.3 Basic Principles

The followings are the illustration of basic principles of trusted AN:

- Accountability requires AN and its provider(s) or vendor(s) to explain, justify and take responsibility for any decision and action made by the AN.
- Equitability for AN and its provider(s) or vendor(s) should take deliberate steps — in the AN life-cycle — to avoid intended or unintended bias and unfairness that would inadvertently cause any harm, damage or loss.
- Explainability is the ability to describe how AN work, i.e., make decisions and actions. Explanations should be produced regarding both the procedures followed by the AN (i.e., their inputs, methods, models, algorithms and outputs, etc.) and the specific decisions and actions

that are made. These explanations should be accessible to people with varying degrees of expertise and capabilities including the public.

Note 1 - For the explainability principle to take effect, the AN engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AN systems, including the ability to explain the sources and triggers for decisions through transparent, traceable processes and auditable methodologies, data sources, and design procedure and documentation.

- Robustness refers to the stability, resilience, adaptability, recency and performance of the systems and machines dealing with changing ecosystems. AN should function robustly throughout its life cycle and potential risks should be continually assessed and managed.
- Safety of AN should be tested and assured across the entire life cycle within an explicit and well-defined domain of use. In addition, any AN should be designed to also safeguard the data, infrastructures, relevant hardwares and softwares which are impacted.

In the above basic principles, none of them is with higher priority than any other, and each of them are related to each other.

12.1.2 General Process of trusted AN

Trust can be divided into the objective part and the subject part. For AN, trust should be considered the subjective part from trustor's perspective and the objective part from trustee's perspective. For trusted AN, both initiation of trust and continuation of trust are the essential factors for networks working and evolving normally. The general process of trusted AN has been illustrated in following Figure 12-2.

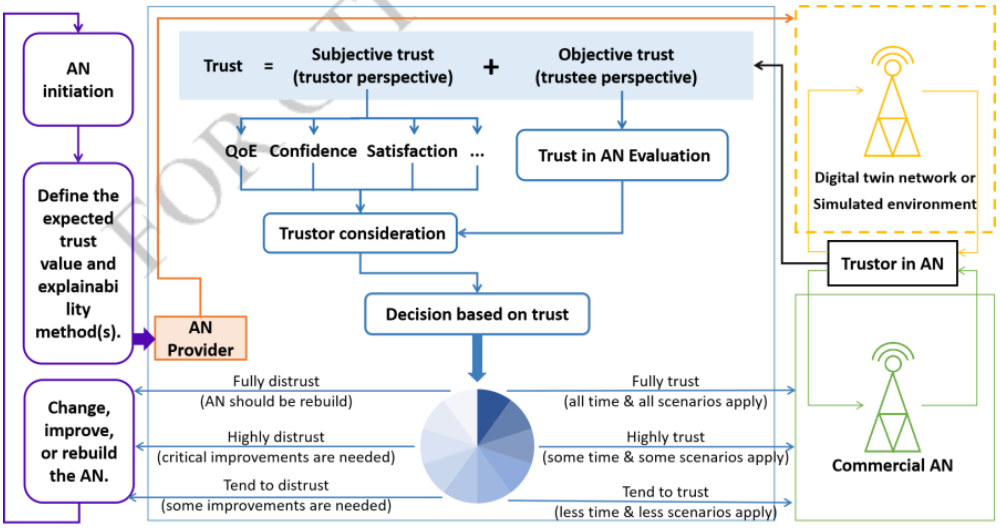


Figure 12-2 General process of trusted AN

12.1.2.1 Initiation of Trust for Trusted AN

In order to initiate trust for trusted AN, trustor in AN can give an order to start the evaluation of trust in AN and the relevant process, after trust in AN evaluation, trustor in AN should take the

results into consideration to make decision to AN.

- Trustor in AN should give some order or process to start the evaluation of trust in AN for the trustee in AN.
- The result of trust in AN evaluation is recommended to outcome as percentage.
- Trustor in AN shall take the result of trust in AN evaluation into consideration, alone with the subjective factors of trust, to make the following authorization decision(s):
  - Fully trust: the commercial AN can be applied in all the scenarios all the time before next evaluation.
  - Highly trust: the commercial AN can be applied in some scenarios in some specific time before next evaluation.
  - Tend to trust: the commercial AN can be applied just for a few scenarios in some specific time before next evaluation.
  - Tend to distrust: the AN provider(s) should make some improvement(s) for current AN, in order to gain more trust in following evaluation(s) .
  - Highly distrust: the AN provider(s) should make some critical and essential improvements for current AN, in order to gain more trust in following evaluation(s) .
  - Fully distrust: the AN provider(s) should rebuild current AN, in order to regain trust in following evaluation(s).
- The evaluation of trust in AN is recommended to take place in the digital twin network or some simulated environment which are both mirrored from commercial AN.

Trusted AN can be initiate when trustor in AN decide to trust, and then the trustee in AN will authorized to govern themselves with minimal to no human intervention.

12.1.2.2 Continuous Trust for Trusted AN

After trusted AN being initiated, continuous or periodic evaluation shall be taken place to monitor the trust's fluctuation, in order to provide reference to trustor in AN to adjust the authorization decisions.

The AN provider(s) should continuously improve the AN, in the meantime, AN may do evolve themselves, so that to earn continuous trust to maintain trusted AN.

12.2 Life Cycle Management

The telecom industry today has a robust and standardized way to handle the Life Cycle Management (LCM) of classical or no-AI based software. This will need to be adapted to handle the new requirements that AI/ML-based software have, such as training and retraining, model-

concept drift, and data requirements. The current industry LCM is not properly prepared to handle this in an effective way. In turn the expected benefits of AI/ML capabilities may not be actualized unless LCM is addressed. The division of stakeholder responsibility, accountability, and deliverables must also be considered and will be impacted by the methods used to meet previously mentioned requirements.

Trust in AI is another concept that will have to be addressed. Since LCM is expected to be able to handle model-concept drift some portion of trust in AI can be addressed by LCM in so far as that deviations from expected behavior can be monitored with a proper LCM setup.

The LCM contains development, validation, operating and retiring software. In this cycle training considerations will raise the need for deciding when and where to use certain datasets, including CSP network data and vendor R&D data. Retraining makes the LCM more iterative as retraining essentially creates a new software version that must then be validated.

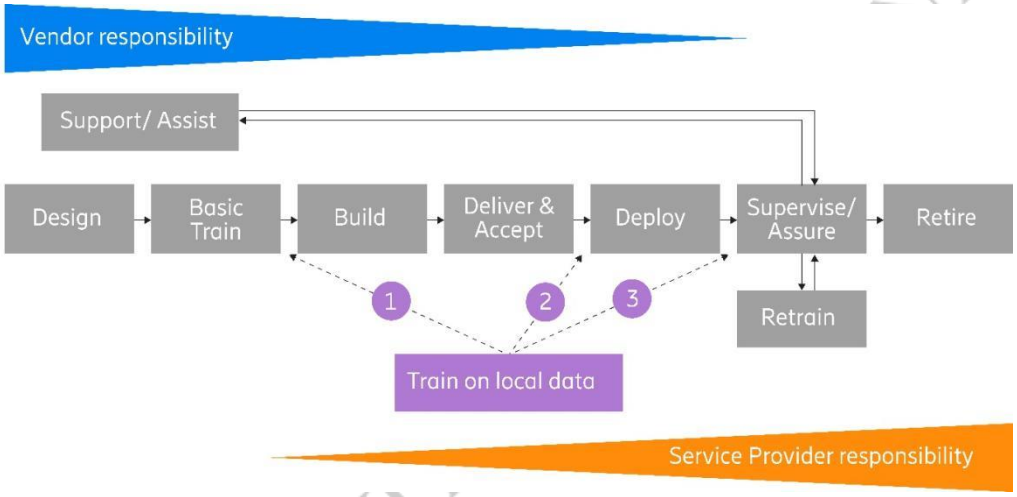


Figure 12-3 AI/ML LCM with options for training with network data

When developing an AI/ML based feature a vendor can use it's R&D dataset to do basic training but in most cases CSP network data will also be required. This can be made available to the model through three main options as shown in Figure 12-3 **错误!未找到引用源。**.

The first option is to make CSP data available to the vendor in their R&D domain, this lets the vendor deliver a fully developed model to the CSP. The second option is to deliver a partially developed model to the CSP who does the final training in their domain. The third is to deploy the model and allow online training for the final adjustments.

All options listed above comes with tradeoffs in accountability, responsibility, deliverables and data access between the CSP and vendor. The options will also require different capabilities like, training and assurance on the part of the CSP.

Standardization will be required to create solutions that can benefit the industry and create a common framework on which to build these.

13 Conclusions and Recommendation

- Autonomous network becomes an important enabler for 5G and future network evolution and enhancement.
- SDOs and industry parties are recommended to work together on network automation enhancement and network intelligence to make progress in a coordinated way.
- The use cases of autonomous network can be categorized based on the dimension of full life cycle and main functional entities.
- Full autonomous network will be a long-term target. Autonomous network level describes the level of autonomy capabilities of the network and it is beneficial for the industry on roadmap planning. Currently, most of the use cases are evaluated between level 2 and level 3.
- Regarding to the autonomous O&M scenarios, the essential capacities of autonomous network have been refined into four categories, i.e., autonomous perception, autonomous diagnosis, autonomous prediction, and autonomous control. AI technologies are gradually being applied in all the processed of the whole O&M life cycle.
- Trust in autonomous network and AI/ML life cycle management are introduced and call for further study to promote the development and maturity of autonomous network technologies and corresponding industries.