GTI 5G Radio Network Intelligence Technical Requirements White Paper

GTI 5G Radio Network Intelligence Technical Requirements White Paper

Version:	V1.2
Deliverable Type	☐ **Procedural Document** √ **Working Document**
Confidential Level	√ **Open to GTI Operator Members** √ **Open to GTI Partners** ☐ **Open to Public**
Working Group	network WG
Task	Integration Across Areas-Autonomous Network
Source members	CMCC, Huawei, ZTE, CICT Mobile, Ericsson
Support members	Nokia Shanghai Bell, H3C, Ruijie Networks, Comba
Editor	CMCC: Jinghua Kuang, Songhe Lu, Lin Zhu Huawei: Yuanxian Tang,Yang Zheng, Zhijie Xu,Chengyu Zhang, Jianfeng Chen, Yan Sun, Xiaoze Xu, Li Wang, Qinxin Liu, Jing Liang, Xianming Li, Jiaming Zhang, Chuan Ma, Junjie Zhang, Jian Fan, Yanli Zhuang, Ying Xu, Baokun Wang, Kunlun Wang, Keke Qiu, Jian Shi, Xi Luo, Xiaowei Qin, Xi He. ZTE: Yong Zhan, Pan Li, Hongtao Zhang, Cheng Guo, Keke Su CICT Mobile: Leming Fang, Lewei Hu Ericsson: Vivian Bai, Rui Wan, Xiaoqing Ma

Last Edit Date	02-09-2023
Approval Date	DD-MM-2023

Document History

Date	Meeting #	Version #	Revision Contents
31-08-2023		V1.0	Initial draft
02-09-2023		V1.1	Adding UC on radio network O&M
04-09-2023		V1.2	Editorial Change

Table of Contents

Abbreviations

Abbreviation	Explanation
3GPP	3rd Generation Partnership Project
5G	5th Generation
5QI	5G QoS Identifier
ADN	autonomous driving network
AGI	artificial general intelligence
AI	artificial intelligence
ANP	access network provider
AoA	angle of arrival
API	application programming interface
CCSA	China Communications Standards Association
CEU	cell edge user
CHR	call history record
CSI	channel state information
CV	computer vision
E2E	end to end
ETSI	European Telecommunications Standards Institute
FDD	frequency division duplex
FDR	flow data record
GHz	Giga Hertz (GHz)
GPS	Global Positioning System
GPT	Generative Pre-trained Transformer
IAB	integrated access and backhaul
ICT	information and communications technology
CSC	communication service customer
CSP	communication service provider
NOP	network operator

IoE	Internet of Everything
IoV	Internet of Vehicles
ISAC	Integrated Sensing and Communications
KPI	key performance indicator
LLM	large language model
LOS	line of sight
MAC	Media Access Control
MIIT	Ministry of Industry and Information Technology
ML	machine learning
MML	man-machine language
MnS	management service
MR	measurement report
SDO	standards developing organization
NaaS	network as a service
NE	network element
NLOS	non-line of sight
NLP	natural language processing
NMS	network management system
NR	New Radio
OMC	operation maintenance center
OPEX	operational expenditure
PLC	programmable logic controller
QoE	quality of experience
RAN	radio access network
RAT	radio access technology
RF	radio frequency
RRM	radio resource management
RSRP	reference signal received power

SLA	Service Level Agreement
TDD	time division duplex
TDOA	time difference of arrival
TOA	time of arrival
TTI	transmission time interval
AV	unmanned aerial vehicle
UC	use case
UE	user equipment
XR	extended reality

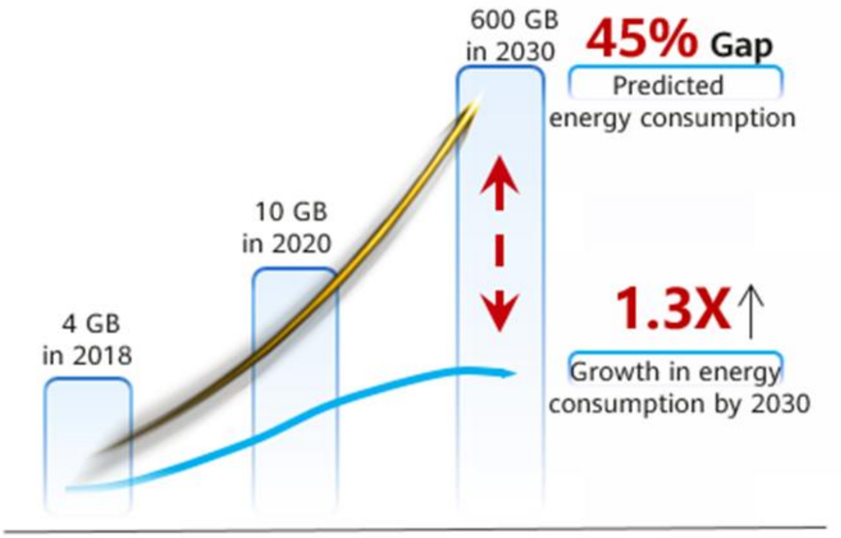
1 CURRENT STATE AND VISION FOR WIRELESS NETWORKS

1.1 New Challenges

The Fourth Industrial Revolution is rapidly ushering in a world where everything is connected and intelligent. All industries are actively exploring the path to digital transformation. Communications networks have shifted their focus from serving individuals to serving the entire digital society. The emergence of the digital economy has gradually become the second evolutionary curve for operators, bringing both new opportunities and challenges for ICT investments in the telecommunications industry.

1.1.1 Growing Wireless Traffic and Energy Consumption

Global operators are prioritizing green and low-carbon strategies as critical to their sustainable development. To date, 32 operators have successfully developed dual-carbon action plans. Energy operating revenues (OPEX) continue to rise as a percentage of operators' revenues. One European operator, for example, has seen a 10% increase in energy consumption over the past three years and has consistently exceeded its 1% target over the past five years. As a result, cost reduction has become a key focus for them. In China, operators have intensified their evaluation of energy consumption. For example, one Chinese operator requires a 1% to 3% reduction in power consumption per logical site and increases the weight of operational efficiency (self-intelligence/cost/energy consumption) from 5% to 20%. In Europe, the escalating energy crisis has led to a sharp rise in electricity costs, with some countries and regions facing energy shortages. Operators are now facing the pressure of rising OPEX and potential service disruptions due to load shedding. Another European operator has outlined a clear strategy to reduce power costs for 35,000 sites on the live network through energy conservation measures while ensuring service continuity.
With the rapid development of services, users are becoming increasingly reliant on mobile networks in their daily lives. It is estimated that mobile network traffic will increase 100-fold by 2030, resulting in the expansion of sites, spectrum and channels. As a result, the energy consumption of wireless networks will increase proportionally. Operators need to balance service development with energy consumption. They are faced with the task of achieving optimal user experience and energy efficiency within complex network and service scenarios.



To meet this challenge, operators must take effective measures to improve energy efficiency.

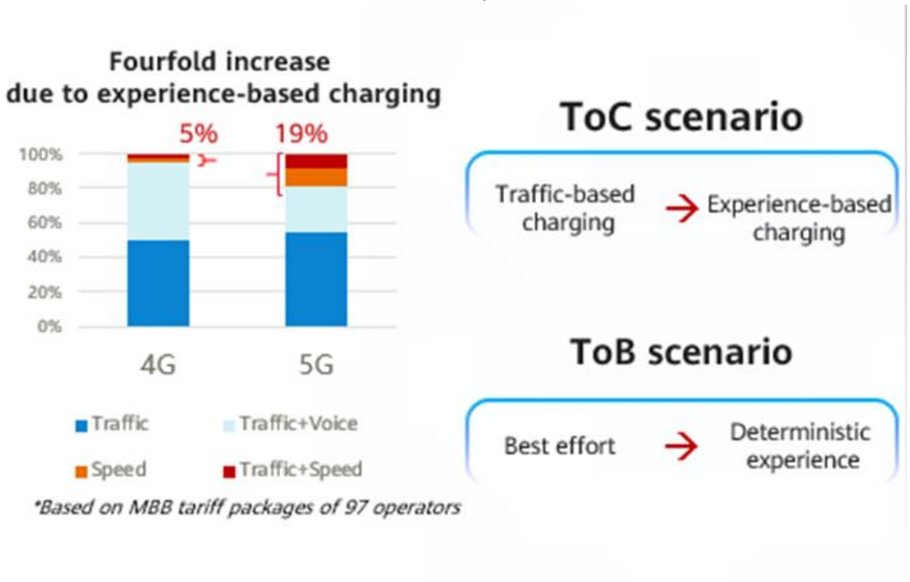
1) Deploy efficient power modules and intelligent cooling systems to reduce base station power consumption.

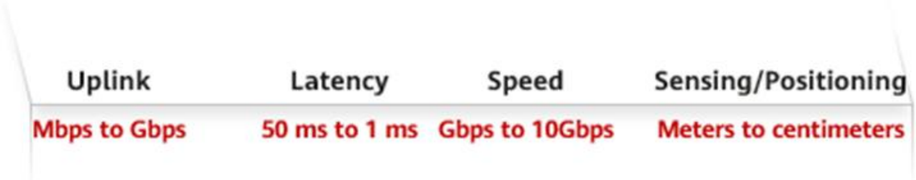
2) Optimize network architecture to reduce the number of base stations.

3) Establish a green O&M system to improve energy management. Establish green O&M standards and processes to control energy consumption and achieve green operations.

1.1.2 Growing Demand for Deterministic Experience as New Services Are Introduced

As commercial network deployments accelerate around the world, wireless networks are gradually transforming from consumer-centric mobile broadband networks to Internet of Everything (IoE). In the upcoming IoE era, both people and things will be connected through wireless cellular networks. The diversification of services brings different network requirements. For example, the immersive Extended Reality (XR) experience requires symmetric uplink and downlink bandwidth, 10 ms latency, and a consistent experience anytime, anywhere; Intelligent Internet of Vehicles (IoV) requires ultra-wide coverage and low-latency control; Unmanned Aerial Vehicle (UAV) services require high-bandwidth uplink backhaul and low-latency downlink control; Industrial Machine Vision relies on 1 Gbps to 10 Gbps uplink bandwidth along with ultra-low latency of 4 milliseconds at 99.999% reliability during the core production phase. In transport and low-altitude detection scenarios, wireless networks are required to provide Integrated Sensing and Communications (ISAC) capabilities. Service diversification also introduces different network coverage requirements. For example, UAV services require network coverage at low altitudes, IoV requires ultra-wide coverage, and industries require coverage in mining areas, oceans, and industrial production environments. In addition, the development of diversified services places different demands on network service level agreements (SLAs). For example, XR immersive experiences and industrial programmable logic controllers (PLCs) require deterministic assurance of network latency.



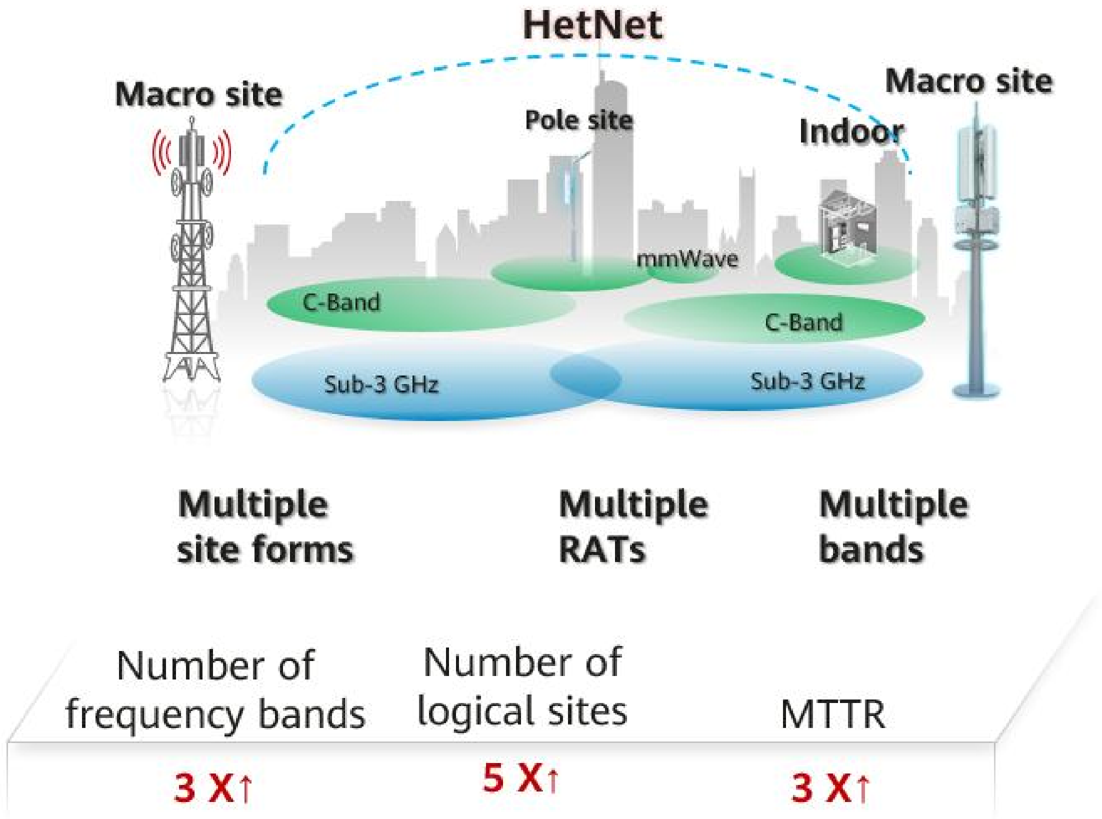


Therefore, 5G networks must meet the following requirements to support a variety of services:

1) Multi-dimensional network capabilities: The network should move from being primarily focused on downlink services to providing diverse capabilities such as uplink, downlink, latency, positioning, reliability, and sensing.
2) Diversified network coverage: The network should evolve from a ground network focused on human users to a multi-dimensional network that provides ultra-wide geographic coverage, low-altitude coverage, and super-distance coverage in oceans, as well as campus and ground coverage.
3) Differentiated SLA levels: The network should provide differentiated, deterministic assurance, rather than just best-effort services.

1.1.3 More Complex O&M Due to the Introduction of New Sites, RATs, and Frequencies

The number of frequency bands increases significantly to more than 15 frequency bands. These bands range from sub-1 GHz and sub-3 GHz to C-band, sub-10 GHz and mmWave. Wireless networks are evolving into FDD+TDD convergence networks. To meet the requirements of various scenarios, the structure of wireless networks will evolve into heterogeneous networks that may consist of macro base stations, mast sites, micro base stations, and integrated access and backhaul (IAB).



To meet this challenge, operators need to incorporate AI technologies to improve O&M efficiency and provide proactive prediction and prevention based on the concept of achieving "zero fault" and maintaining always-online wireless networks.

1) For massive alarms, intelligent alarm identification and diagnosis is required to accurately locate the root cause of alarms and help operators accurately troubleshoot.
2) Fault prediction and prevention can help operators proactively identify network risks in advance, enabling proactive O&M and facilitating the establishment of "zero fault" wireless networks.

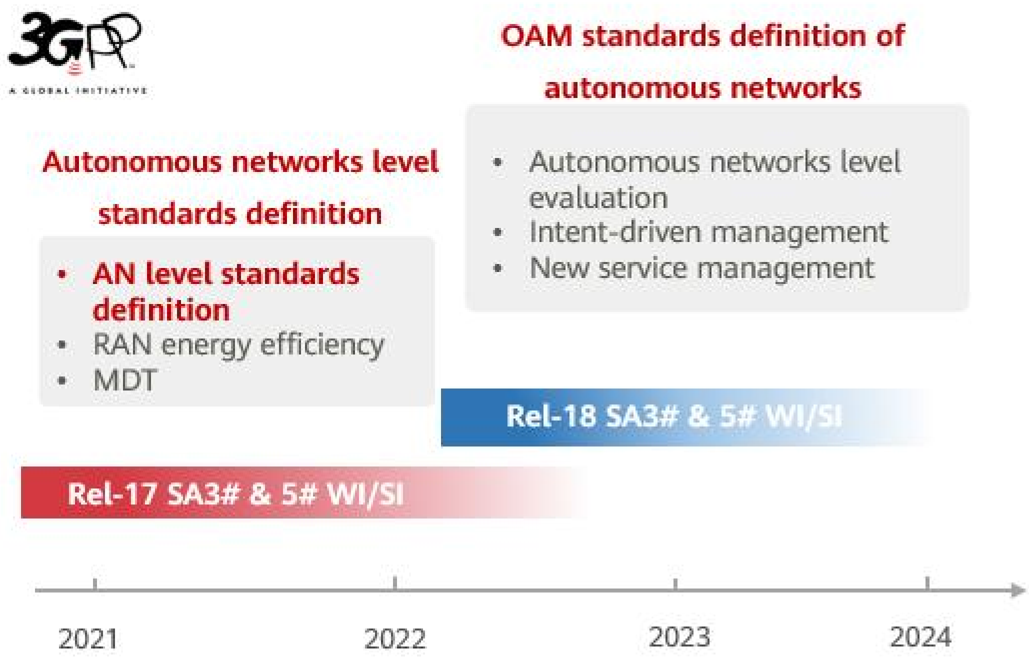
1.2 Concept and Status Quo of Autonomous Networks

Autonomous networks enable networks to become more automated and intelligent. By incorporating automation and intelligence technologies such as AI, these networks aim to achieve predictability and operational autonomy, enabling the development of automated and intelligent O&M capabilities across the entire communications network lifecycle.

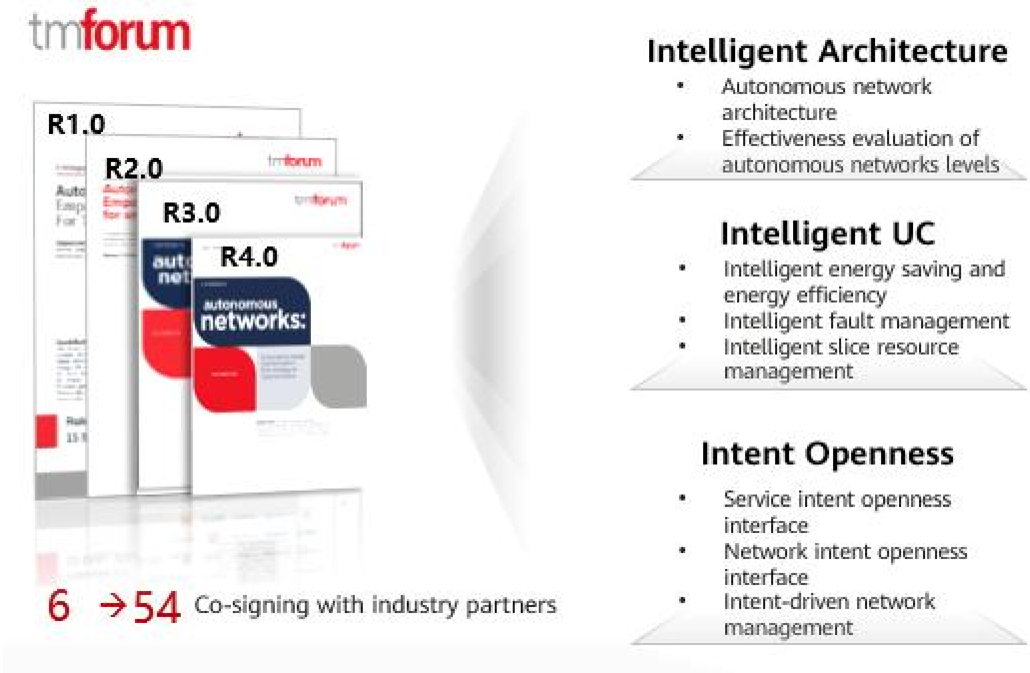
1.2.1 Evolution of Autonomous Networking Industry Standards

With the diversification and rapid development of new mobile network application services, operators have raised their expectations for ensuring service quality and improving network efficiency. They have taken measures to promote the adoption of technologies such as AI in the telecom industry. In addition, the rapid growth of mobile traffic has placed new demands on spectral efficiency and the development and optimization of new network technologies. Operators are facing significant challenges in their digital transformation journey, including meeting diverse service demands, improving customer experience, streamlining upstream and downstream collaboration, and improving collaboration efficiency in a multi-vendor environment. The industry has reached a preliminary consensus that network intelligence and automation are key technical means to address these emerging requirements and challenges. As technology frontrunners, standardization and industry organizations in and outside China, such as the 3rd Generation Partnership Project (3GPP), TM Forum, European Telecommunications Standards Institute (ETSI), and China Communications Standards Association (CCSA), have already initiated standardization efforts in their respective technical areas. They are actively exchanging technologies between standards organizations through joint conferences and contact letters. The participation of standards organizations in enterprise discussions can identify the optimal path for intelligence evolution, help operators collaborate effectively with partners in digital network transformation, streamline end-to-end processes in multi-vendor environments, and enhance competitiveness in rapid service deployment.

3GPP is an international standards organization responsible for developing standards for mobile communications technologies. The standards and specifications developed by 3GPP are managed through a series of releases, typically formulated every one to two years. 3GPP Releases 18, 19, and 20 represent the second innovation of 5G, known as 5G-Advanced. New features in 3GPP Release 18 include the integration of AI and machine learning technologies, which provide data-driven and intelligent network solutions. RAN intelligence has become a major research project within 3GPP Release 18. Releases 19 and 20 are currently under active development, with a focus on advancing autonomous networks in the industry.



TM Forum launched the Autonomous Networks Project (ANP) in May 2019. The project aims to define fully automated, zero-wait, zero-touch, zero-trouble innovative network/ICT services for users and consumers in vertical industries. In addition, TM Forum has taken the lead in organizing multi-SDO work across multiple standards organizations to achieve consensus on concepts, frameworks, and key viewpoints in the autonomous networks domain while fostering cross-organizational collaboration. To date, TM Forum has published several specifications related to autonomous networks, covering architecture, evolution levels, intelligence use cases (UCs), and intent openness. Through its Autonomous Networks Collaboration Project, the organization actively supports the development of wireless intelligence and provides industry guidance.



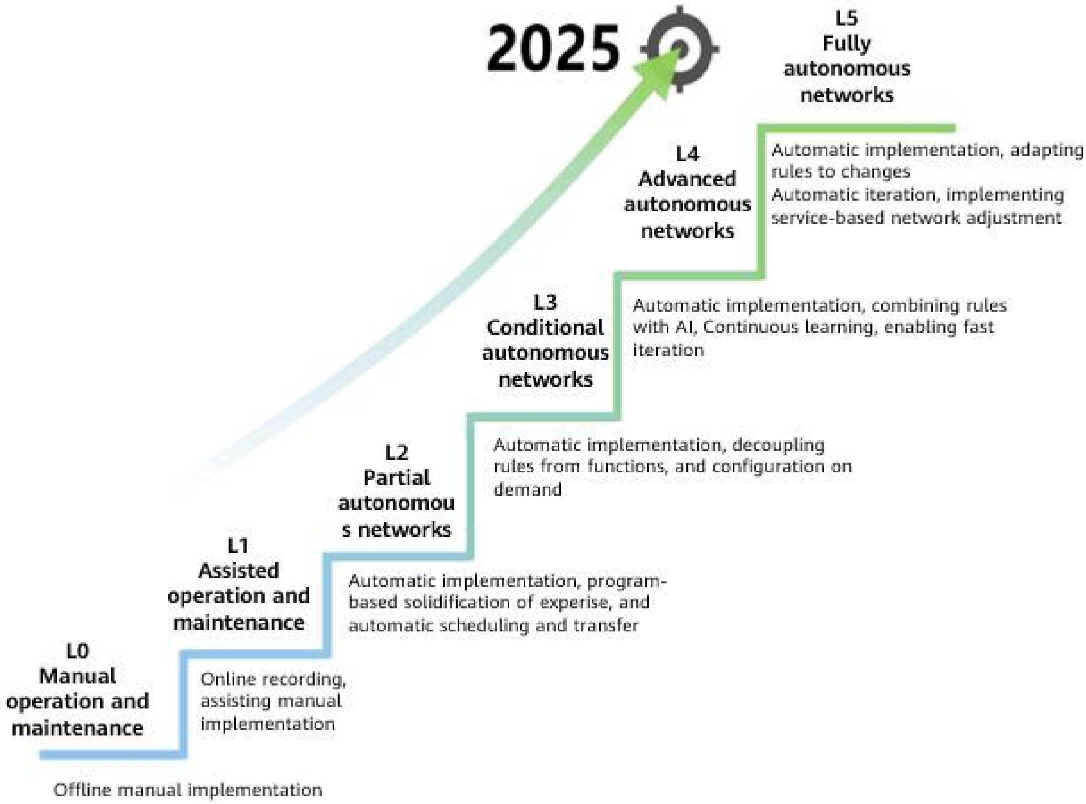
ETSI is an independent, nonprofit standards organization focused on the development of information and communications standards. ETSI has recently published a white paper entitled "Unlocking Digital Transformation with Autonomous Networks", which highlights the importance of autonomous networks in achieving digital transformation. This white paper serves as a proactive initiative by ETSI to promote the advancement of wireless intelligence and provide guidance to the industry.

On September 23, 2021, TM Forum and CCSA co-hosted the 2021 Autonomous Networks Industry Summit in Beijing to strengthen consensus within the autonomous networks industry and share the latest progress in standards development and business achievements. During the event, key figures from China's Ministry of Industry and Information Technology (MIIT), CCSA, and TM Forum jointly announced three industry promotion initiatives, including clarifying the generational characteristics of autonomous networks and guiding their development direction; improving the working mechanism and establishing industry standards; and promoting initial pilot projects and accelerating the deployment of autonomous networks in live networks. CCSA is actively promoting the development of RAN intelligence and providing guidance to the industry through the Summit.

In summary, a global consensus has been reached among operators on the deployment of autonomous networks, with L4 autonomous networks as the phase target for operators. By the year 2022, more than 10 leading operators worldwide had announced their commitment to achieve L4 autonomous networks by 2025. These networks have become an integral tool for operators to transform their data intelligence and are currently undergoing rapid development.

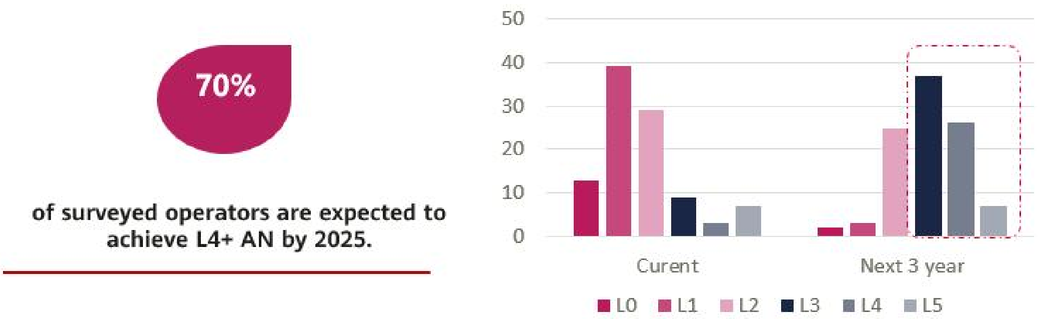
1.2.2 Autonomous Networking Practices

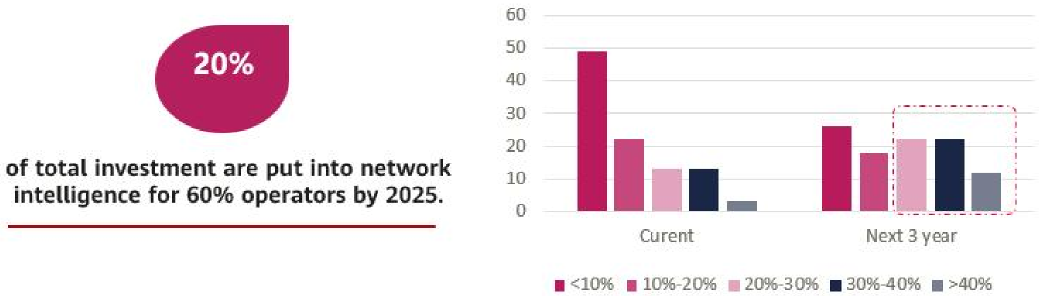
The Autonomous Networks Framework divides network autonomy capabilities into six levels: L0 to L5. The Autonomous Networks levels from L0 to L5 indicate different network features and capabilities. The Autonomous Networks Framework, in accordance with TM Forum's Guiding Principles, outlines the characteristics of each Autonomous Networks level as it relates to the implementation of IT systems, taking into account the practical needs of network operations and maintenance.



China Mobile has been at the forefront of the autonomous networks industry, leading in both development and implementation. With the concept of autonomous networks in mind, the company plans to digitally and intelligently transform network O&M, strengthen the building of its automation and intelligence capabilities, and set the overall goal of achieving L4 autonomous networks by 2025. China Mobile is building intelligent infrastructure, agile operations and all-scenario services through fully automated networks and ICT to provide optimal customer experience for vertical industries and consumers. In addition, the operator actively promotes innovation in the theory of autonomous networks, develops capabilities, and implements large-scale applications. With a focus on supporting customer development and consolidating quality leadership, China Mobile is continuously optimizing three closed-loop processes in terms of customer requirements, end-to-end services and professional networks. The company is also committed to building and improving the intelligent operation and maintenance capability of networks and network management systems.

To achieve the goal of achieving L4 autonomous networks by 2025, China Mobile is actively exploring the ecosystem cooperation system, industry enablement platform, and new O&M service mode. In addition, the company is continuously deepening industry cooperation to promote the digital and intelligent transformation of O&M, ultimately achieving industry development and prosperity through joint efforts. In the coming years, China Mobile will continue to focus on building autonomous networks, accelerating the digital and intelligent transformation and upgrading of O&M, and improving network quality. These efforts are not only essential to promote value operations on a large scale, but also imperative to facilitating the transition from relying solely on demographic dividends to embracing intelligence dividends. With the ongoing development and implementation of AI technologies, we believe that autonomous networks will bring us more surprising experiences in the near future.





Source: <autonomous networks: from concept to reality> by TM Forum,2022

With the continuous advancement of AI technologies, more and more operators are realizing the importance of network intelligence in improving the efficiency of network operations and delivering better service experiences to customers. As a result, many operators are proactively investing in AI technologies to improve their competitive edge.

By 2025, 70% of operators in the global telecom industry are expected to invest in developing AI software, hardware, and services, and building intelligent networks, including network planning, O&M, optimization, and customer services. 60% of operators will invest in AI technologies to implement network intelligence, accounting for more than 20% of their total investment. With the continuous development and implementation of AI technologies, we believe that network intelligence will bring us more surprising experiences in the future.

1.2.3 Outlook of Autonomous Networks

Native intelligence is one of the visions of network autonomy. It means that communications networks have full AI capabilities, including data, computing power, and algorithm capabilities. The development of this technology will enable communications networks to better meet the functional and logical requirements of intelligent applications in the future.

The white paper *6G Vision and Candidate Technologies* published by the China Academy of Information and Communications Technology also mentions that new air interfaces and new network architectures with native intelligence are among the ten candidate technologies of 6G. With these technologies, network performance will be greatly enhanced, meeting the requirements of new services and scenarios in the future, serving the intelligent society and life, and contributing to the realization of the 6G visions of "intelligent connection of everything and digital twin."

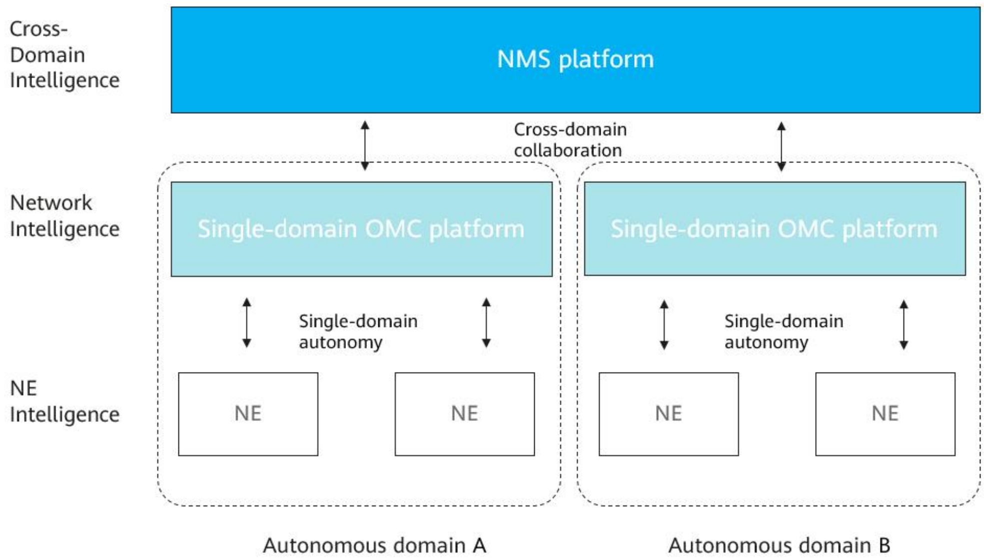
In the future, autonomous network technologies will continue to evolve toward intelligent scenarios and native intelligence. Intelligent scenarios: The closed-loop intelligent system with visible benefits, perceptible scenarios, plastic policies, and controllable effects is built to implement efficient and accurate scenario-based services and continuous evolution. Native intelligence: The integrated computing-network architecture integrates AI computing capabilities into communication processes, enhancing the intelligent learning capability and scenario adaptability for the physical layer, MAC layer, and network layer, and enabling continuous evolution of personalized intelligent service capabilities. It also integrates the intelligence layer and data layer into the network architecture design, NEs, and interface implementation. As a result, the future intelligent network can internally implement self-adaptation, self-learning, self-correction, and self-optimization.

In a word, native intelligence, as one of the future visions of network autonomy, will play an important role in the communications field. It will drive continuous innovation in communication technologies, promote industry transformation and upgrading, and contribute to the sustainable development of human society.

2 Technical Requirements for RAN Intelligence

2.1 RAN Intelligence Architecture

The RAN intelligence architecture consists of three layers. It addresses the challenges of the network fabric. The network management system (NMS) platform is responsible for managing intelligence across domains and vendors. The single-domain operation and maintenance center (OMC) platform is responsible for single-domain wireless network intelligence. The base station layer provides single-site NE intelligence. In this way, layered processing can make the entire network more efficient and intelligent.



Cross-Domain Intelligence: Cross-domain and cross-vendor network-level intelligence can globally monitor and predict the health of the entire network by aggregating and analyzing data from different domains and vendors. In addition, collaborative algorithms can be used to optimize resource allocation across domains and vendors, improving overall system performance and efficiency.

Network Intelligence: Network status can be monitored and predicted through centralized management and data analysis of wireless base stations. In addition, intelligent algorithms can be used to optimize resource allocation at the network level, improving network performance and efficiency.

NE Intelligence: Base stations can use embedded intelligent technologies to implement real-time perceiving, modeling, prediction, and multi-dimensional decision making. Intelligent algorithms can be used to implement on-demand resource configuration, provide best experience, and achieve optimal capacity to achieve optimal performance and energy savings.

In a word, the architecture of RAN intelligence is evolved according to network challenges. The goal is to implement more efficient intelligence through layered processing.

2.2 Technical Requirements for RAN Intelligence

As wireless networks evolve, there are technical bottlenecks in the operation and optimization of different network layers. As a result, networks and NEs face enormous challenges in terms of perceiving, analysis and decision making, intent, and experience. To meet these challenges, operators and vendors must continue to innovate and upgrade technologies to improve network intelligence.

Intelligence Capability		Technical Requirements	Technical Direction			
			Intelligent Data perceiving	Digital Twin	Intent Openness	Intelligent Air Interface
Network intelligence (planning, building, maintaining, and optimizing)	Network perceiving	◆ Precise UE positioning ◆ Precise device engineering parameters ◆ Precise environment modeling ◆ Precise radio signal measurement	Y			
	Analysis and decision-making	◆ Multi-objective optimization ◆ Multi-method parallel optimization ◆ One-shot optimization		Y		
	Human machine interaction	◆ Natural language interaction ◆ Simplified interface invocation			Y	
NE intelligence (performance experience)	Wireless perceiving	◆ Efficient L3 multi-frequency measurement	Y			
	Air interface interaction	◆ Efficient L1 channel measurement				Y

In terms of perceiving, operators need to accurately monitor network status in real time, collect and analyze large amounts of data, and use efficient data analysis algorithms and powerful data processing capabilities to accurately model the environment.

In terms of analysis and decision-making, operators need to perform in-depth analysis based on the collected data, identify problems, and formulate multi-objective optimization solutions. This requires digital twins to build a simulation verification environment, and powerful computing capabilities and intelligent decision-making algorithms to develop the one-shot optimization capability.

In terms of intent, operators need to formulate appropriate intent interfaces based on customer requirements and service development, use models to build simplified O&M based on human-machine interaction using natural language.

In terms of experience, operators need to continuously optimize network architecture and upgrade technologies to meet customers' requirements for network speed, stability, and security, and deliver better customer experience.

Based on the practices in recent years, we summarize the technical requirements on network and NE intelligence and explore the technology development to promote the rapid development of RAN intelligence.

2.2.1 Network Perceiving

2.2.1.1 Precise Service Recognition

By embedding an intelligent service recognition and analysis algorithm module in the base station side, it can achieve fine-grained recognition of real-time data services. Taking TikTok as an example, it can finely identify TikTok short videos, TikTok live broadcasts, TikTok web browsing, TikTok calls, TikTok wallet and so on. The intelligent algorithm is based on business profiling to obtain its corresponding feature information, and for different businesses, different training algorithms can be used to obtain relevant information and enrich its feature library.

By customizing the Type type to extend the generalized QoS guarantee strategy, after the service type is identified, different services can be mapped to different Type types and adopt different guarantee strategies. For example, for rate-based services, according to different Type guarantee strategies, the scheduling priority and guaranteed rate are improved; for delay-based services, according to different Type guarantee strategies, the pre-scheduling priority and pre-scheduling frequency are improved. And according to gap of the real-time perception evaluation results and the target, scheduling parameters can be dynamically adjusted(such as AMBR parameters, SBR parameters, Type level, etc.).

Through the self-built service experience quality evaluation xEMI system on the wireless side, the base station real-time evaluates the QoE indicators at the service level, and adjusts the real-time guarantee strategy according to the evaluation results, taking the service experience as the goal to quickly achieve closed-loop optimization; at the same time, combining business application data , Wireless air interface KPI and CDR and other data for multi-dimensional analysis, efficiently locate the root cause of poor wireless perception of business applications, support network-level precise resource management, and achieve deterministic service experience.

2.2.1.2 Precise UE Positioning

Traditionally, sites evaluate user equipment (UE) locations based on terminal channel measurements and positioning algorithms. Currently, single-site positioning and multi-site positioning are supported.

1) Single-site positioning: The time of arrival (TOA) is used to determine the distance between a UE and a cell, which is the radius of a circle. Then the angle of arrival (AoA) and the antenna azimuth are used to determine the angle between the ray and the coordinate system. The intersection of the ray and the circle is the UE position.
2) Multi-site positioning: When UE signals can be measured from multiple base stations, the multiple intersections of the rays and circles can be obtained. Using algorithms to combine multiple sets of results can achieve better accuracy compared to single-site positioning.

Actually, wireless networks are designed for communication rather than positioning. At present, wireless networks face the following challenges

1) Inaccurate measurement: The Radio Resource Management (RRM) measurement of services and broadcast is used instead of the traditional pilot dedicated to positioning. The bandwidth is small and the accuracy is low.
2) Large error in engineering parameters: Networks are built for communication instead of positioning. The maintenance of engineering parameters is poor. The error of engineering parameters on the live network is usually more than 30 degrees.
3) Poor environment: Most UEs are located in densely populated urban areas. More than 90% of UEs are located in one or more Non-Line-of-Sight (NLOS) scenarios, where the signal propagation path is longer, and the error of TOA or time difference of arrival (TDOA) estimation even reaches 200 m to 500 m. In addition, the AoA measurement is affected by scattering and refraction around the UEs and base stations. The error of angle measurement accuracy can reach 20 to 30 degrees.
4) The accuracy of indoor positioning based on UE reporting is low and the cost is high.

2.2.1.3 Precise Device Engineering Parameters

engineers use instruments to measure the engineering parameters, including GPS coordinates, antenna posture, and azimuth onsite.

The traditional solution relies mainly on manual operations. However, measurement results may be inaccurate because of engineers skill differences or instrument problems. Take the azimuth parameter measurement as an example. The error of azimuth measured by the compass is usually greater than 30 degrees due to the precision difference and interference from magnetic materials such as metal towers or equipment.

2.2.1.4 Precise Environment Modeling

Traditionally, environmental intelligence collection devices consisting of sensors such as cameras and radars are used to collect image and point cloud data in the field. The data is then used to generate scenario reproduction maps of the collection area based on environment reconstruction algorithms. In this way, the environmental information of the area can be obtained. This traditional mode has the following drawbacks

1) Devices need to be carried manually to collect data, resulting in high labor costs. In addition, sensors need to collect data from all locations in the environment, which is time-consuming in complex environments such as large factories.
2) Due to factors such as the cost of lidar sensors, high-precision environmental mapping devices in the market are expensive and difficult to promote on a large scale.
3) The generated maps cannot be updated. They become invalid when the environment changes. The data must be collected again.

The intelligent solution collects massive wireless data generated in communication services to model the environment and detect environmental changes. In this way, the wireless and physical environments can be reconstructed at low cost.

2.2.1.5 Precise Radio Signal Measurement

As wireless networks continue to evolve, more frequencies are being used for wireless applications, making wireless networking more complex. In complex multi-frequency networking scenarios, it is crucial to efficiently and accurately obtain multi-frequency networking information during inter-frequency handovers or carrier selection to take full advantage of multi-frequency networking and provide an optimal user experience.

A UE performs inter-frequency handovers using inter-frequency measurements and selects an inter-frequency cell to camp on based on the measurement results. During the carrier selection, the base station evaluates the RSRP value reported by the UE to decide whether or not to change the carrier.

The traditional approach to selecting the optimal service carrier for a UE is limited by a fixed handover threshold and cannot prioritize carriers based on optimal user experience. In addition, the handover process using the traditional inter-frequency measurement method introduces a delay of approximately 300 ms. This delay prevents the UE from promptly switching to the optimal service carrier, resulting in a decrease in the user-perceived rate and inability to support delay-sensitive services.

2.2.2 Analysis and Decision-Making

In traditional automatic O&M, AI has been applied to various application functions, including traffic prediction, abnormal KPI detection, fault identification, and energy saving based on tidal traffic characteristics, and has achieved a certain level of intelligence. As autonomous networks evolve toward L4, higher-level intelligent technologies are required.

2.2.2.1 Multi-Objective Optimization

Wireless network evaluation KPIs include coverage, access, call drop, mobility, experience rate, edge rate, traffic volume, delay, utilization and power consumption. In the existing KPI evaluation system, the status of wireless network KPIs can be displayed. When one or more KPIs deteriorate, or multiple KPIs need to be optimized to meet the requirements of new services, the serial mode optimization is traditionally used for the KPIs.

In traditional mode, only one target in one region can be optimized at a time.

When the second KPI is optimized, the optimization parameters may overlap. As a result, the network may be degraded. For example, energy saving optimization may cause the Cell Coverage rate or access rate to deteriorate. You cannot optimize for two or more KPIs.

2.2.2.2 Multi-Method Parallel Optimization

A wireless network contains many network optimization parameters, including cell coverage parameters, inter-cell coordination parameters, and intra-cell optimization parameters. Currently, a network parameter optimization method focuses on a single object, and only one object can be optimized at a time. For example, cell coverage and inter-cell mobility cannot be optimized at the same time. Multi-method parallel optimization cannot be implemented.
In the conventional solution, optimization methods are used in serial mode. For example, coverage optimization, capacity optimization and energy saving are performed one after another. The optimization period sometimes exceeds one month. Parallel optimization cannot be performed because the optimization methods affect each other.

2.2.2.3 One-shot Optimization

With the development of wireless networks, the wireless network structure is becoming more complex, including multiple radio access technologies (RATs), multiple frequencies, multiple antennas, and different site types. It poses great challenges to network O&M. The entire network cannot be accurately described. As a result, iterative optimization may be attempted. This is impractical in some scenarios, such as competition scenarios like the World Cup and Asian Games, where competitions last several hours. Therefore, a single optimization is required.
In the traditional optimization solution, iterative optimization is performed on the live network to obtain the optimal combination of parameters. It may affect the performance of the live network and cannot be applied to scenarios with high timeliness requirements, such as special event scenarios.

2.2.3 Air Interface Interaction

In order for wireless air interfaces to realize strong transmission performance, wireless channels must be accurately characterized and detected, and signal transmission and reception should be adapted to wireless channels. Traditionally, the mathematical basis of information theory is used for theoretical abstraction and mathematical modeling on each wireless transmission module. The universal air interface design based on this method can achieve good robustness and is easy to implement. However, it is difficult

to achieve optimal performance in complex and diversified wireless channel environments.
Consider channel state information (CSI) feedback in an NR network as an example. The codebook-based feedback solution mathematically transforms a measured channel to be fed back into an angle-delay domain and performs parameter quantization on the channel information based on the strong sparsity in the angle-delay domain. In this way, only the angle-delay component with a maximum energy amplitude can be fed back, and the receiver can recover high-precision channel information by inverse transformation while reducing the feedback volume.
In the CSI feedback solution, the angle-delay domain may not be the best transformation domain. The parameter quantization method discards the relatively small amplitude angle-delay components, resulting in a loss of information. As a result, the performance of the conventional CSI feedback solution is much lower than that of the ideal CSI feedback. One of the main directions of current air interface research is to use intelligent algorithms to improve the accuracy of CSI feedback data and approach the upper limit of ideal CSI feedback performance.

2.2.4 Human-Machine Interaction

With the evolution of RATs and frequency bands, diversified wireless network types and heterogeneous networks are emerging. The coexistence of multiple RATs and increasing service complexity pose a great challenge to network operation and maintenance (O&M). However, the openness of network capabilities based on cooperation among network layers can help encapsulate the network complexity and differences of different vendors. Based on open APIs, they are seamlessly connected to operation platforms, network management platforms, and networks based on upper-layer service requirements, achieving end-to-end service closure and operation automation, thereby realizing network self-configuration, self-healing, and self-optimization. Operators provide unified and rich open APIs that can be flexibly invoked by third-party applications, enabling network capability monetization and industry application innovation, thereby building a new NaaS business model.

2.2.4.1 Simplified Interface Invocation

The openness requirements vary according to the stage of the intelligent autonomous network. To achieve higher-level intelligent autonomous networks, the openness of existing basic capabilities must gradually move to intent-based openness. In the intent-based openness mode, O&M engineers, as intent consumers, only need to express their expectations to the network in simplified languages. Intent producers can use knowledge and graph technologies to build a domain knowledge base to translate the expectations into intents. The intents can then be automatically translated into network configuration policies and configuration parameters without manual intervention.

To meet the diverse requirements of upper-layer applications, the traditional domain knowledge-based solution needs to impose strong constraints on the scope of the intents in advance and make deterministic task orchestration assumptions based on the intent translation result to ensure reliable execution of the intents. However, as the network structure becomes more complex, more diverse types of services are carried over the network, and various openness requirements are growing, often exceeding the preset intent scope. In addition, the number of customized service flows increases due to the increased technical difficulty of knowledge accumulation, extraction, and reasoning based on the knowledge graph. Therefore, the request response time is shortened.

2.2.4.2 Natural Language Interaction

In a traditional human-machine interaction approach, basic semantic recognition results are combined with pre-established processing solutions to perform routine tasks.

The traditional approach primarily relies on pre-established processing solutions rather than language models such as GPT. A significant number of solutions must be constructed based on expert knowledge to handle basic network optimization and O&M operations. However, in complex scenarios, the semantic conversion accuracy is low and the system may not provide an appropriate processing solution. As a result, the implemented solution may be significantly different from the actual intent requirements. The development of accurate interaction capabilities based on natural languages is a critical aspect of human-machine interaction.

2.3 Technical Direction of RAN Intelligence

2.3.1 Intelligent Data Perceiving

2.3.1.1 User Perceiving

User perceiving capability, also referred to as user perceiving and positioning capability, is a widely used feature in modern wireless networks. Currently, user perceiving is being developed in two main directions: user location perceiving capability for horizontal and vertical positioning of users, and user environment perceiving capability for identifying the building and street where a user is located. However, there is a need to improve the maritime user perceiving capability and the user status perceiving capability, which include indoor/outdoor user identification and street user identification. Common positioning technologies include proxy positioning, geometric positioning, and fingerprint-based positioning.

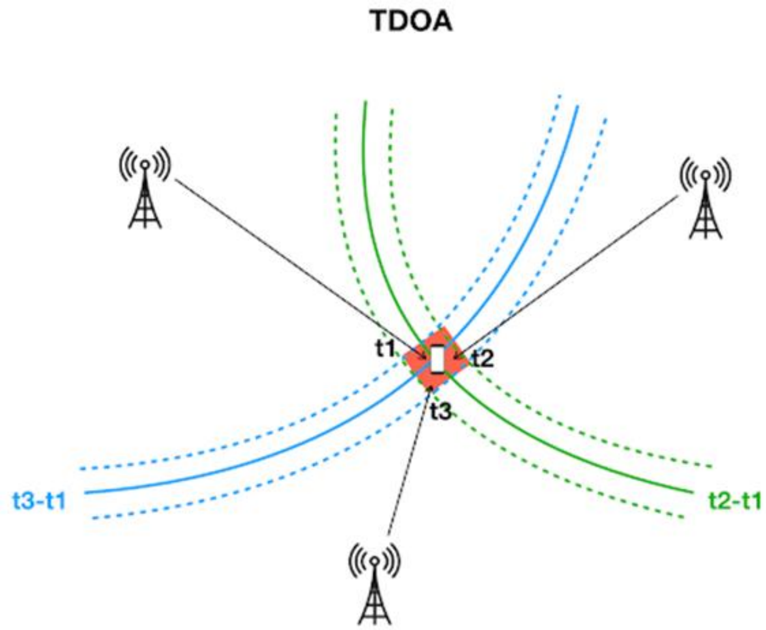
Proxy Positioning

The location of the mobile station can be determined by searching for the closest access point to the user and using the location of the access point or its vicinity as the positioning result. The algorithm is straightforward, but its accuracy is limited by the cell coverage radius. Smaller cell coverage radius results in higher positioning accuracy.

Geometric Positioning

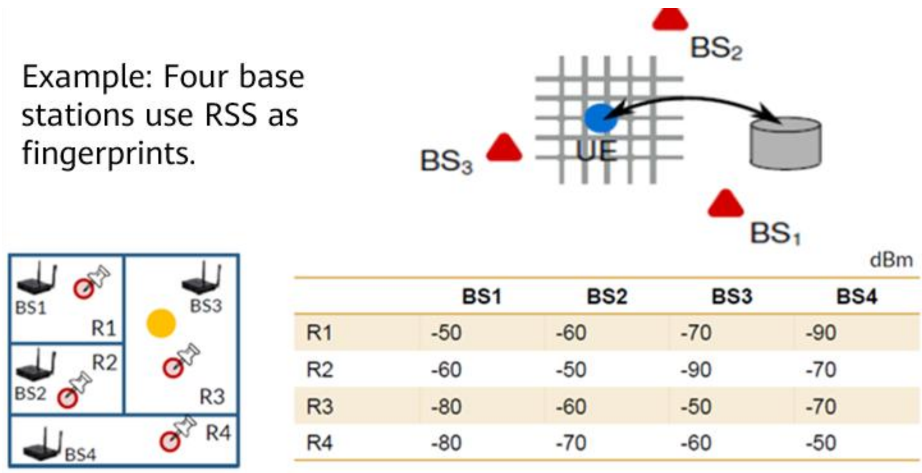
Geometric positioning is divided into multi-site, angle, and time difference of arrival (TDOA) based positioning.

1) Single-site positioning: By calculating the intersection point between the ray and the circle, single-site positioning can determine the location of a mobile station based on the distance and angle between the signal source and the base station. This is a basic positioning method.
2) Trilateration: By measuring the distance between the signal source and each base station, trilateration can determine the location of a mobile station. The base station is the center, and the distance is the radius.
3) Several groups of hyperbolas are constructed based on the distance differences to obtain the intersection point. The TDOA-based positioning algorithm eliminates the effects of varying terminal processing performance, unlike the time of arrival (TOA) based positioning algorithm. When base stations are accurately synchronized, the accuracy of the distance difference estimation depends on the performance of the delay estimation algorithm on the NE side, resulting in improved positioning accuracy.
4) Angle-based positioning: If the base station is equipped with an antenna array, the angle of incidence can be determined from the signals transmitted by the mobile station. The position coordinates of the mobile station can be determined by constructing the intersection of the rays in the direction of the angle of incidence with each base station as the starting point.



Fingerprint-based Positioning

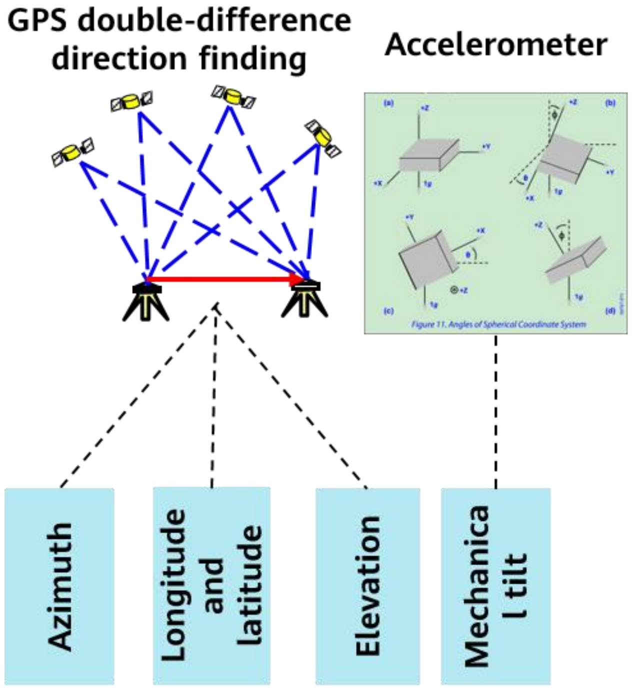
The uniqueness of radio measurements in a given time and space makes it possible to model radio feature data at known locations and perform fingerprint matching on unknown measurements for positioning. The accuracy of fingerprint-based positioning is closely related to feature size and model stability.



2.3.1.2 Engineering Parameter Perceiving

Accurate engineering parameters are fundamental to network optimization and planning. Accurate engineering parameters are essential for calculating target coverage based on traffic distribution. In addition, accurate site azimuth information is required for high-precision user location. To achieve these goals, several aspects need to be optimized and developed.

1) Engineering parameters need to be verified, including antenna azimuth, antenna downtilt, antenna position, and antenna height.
2) Antenna exception detection is required, including antenna blocking, reverse cell connection, cable sequence disorder, and end exception detection.
3) Engineering parameter perceiving involves the use of antenna posture sensors, which use sensor technology to monitor the mechanical tilt, azimuth, elevation, longitude, and latitude of the antenna in real time. This enables network optimization engineers to obtain accurate engineering parameters in time without having to visit the site.

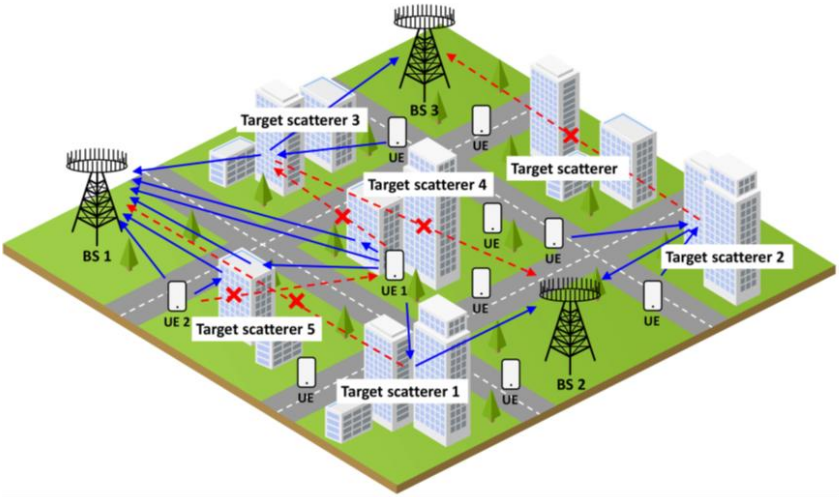


2.3.1.3 Environment Perceiving

To proactively identify and detect potential hazards and risks to wireless networks caused by environmental changes, it is necessary to explore methods for describing, modeling, and sensing virtualized environments. To achieve these goals, several aspects need to be optimized and developed.

1) Environment modeling is a critical aspect that involves reconstructing the physical environment and digitally expressing the radio channel environment, which has no physical meaning but is used to predict the status of the air interface.
2) Interference perceiving is necessary, which involves detecting jammers and pseudo base stations.
3) Intrusion perceiving is also vital, which entails detecting drone activity and identifying any breaches in high-speed rail fences.

A commonly used technology for environment perceiving is the radio propagation environment perceiving algorithm. This algorithm analyzes the channel data of wireless network users to perceive the users, their environment and the network status. It then builds a parameterized physical model with high reliability, which provides significant advantages for accurate system control and performance prediction. It can also implement a comprehensive perceiving of the state of the environment for both the receiver and the transmitter. This includes a range of factors such as the distribution of fixed and dynamic scatterers, the propagation topology between scatterers, the geometric shape and surface structure of objects, and deep material features. By obtaining these feature descriptions, we can effectively create a digital representation of the radio environment, allowing accurate prediction of channel changes. This is important for optimizing communication systems and improving their performance.

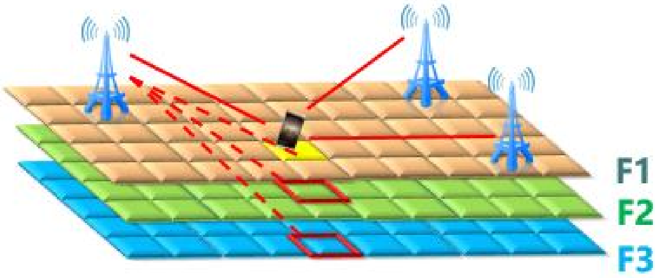


2.3.1.4 Networking Perceiving

As wireless networks continue to evolve, more frequencies are being used for wireless applications, making wireless networking increasingly complex. Efficient and accurate identification of a user's current network environment is critical to fully utilizing the resources of multi-frequency wireless networks and improving the user experience. Virtual Grid technology has been introduced to achieve this goal.

1) This technology uses multi-dimensional measurement of radio signals to classify UEs with similar radio characteristics into the same virtual grid. The virtual grid information is collected and used to build a model of the current network.
2) This model is then used to construct a comprehensive radio network scenario that provides reference information for multi-frequency handovers and carrier selection.

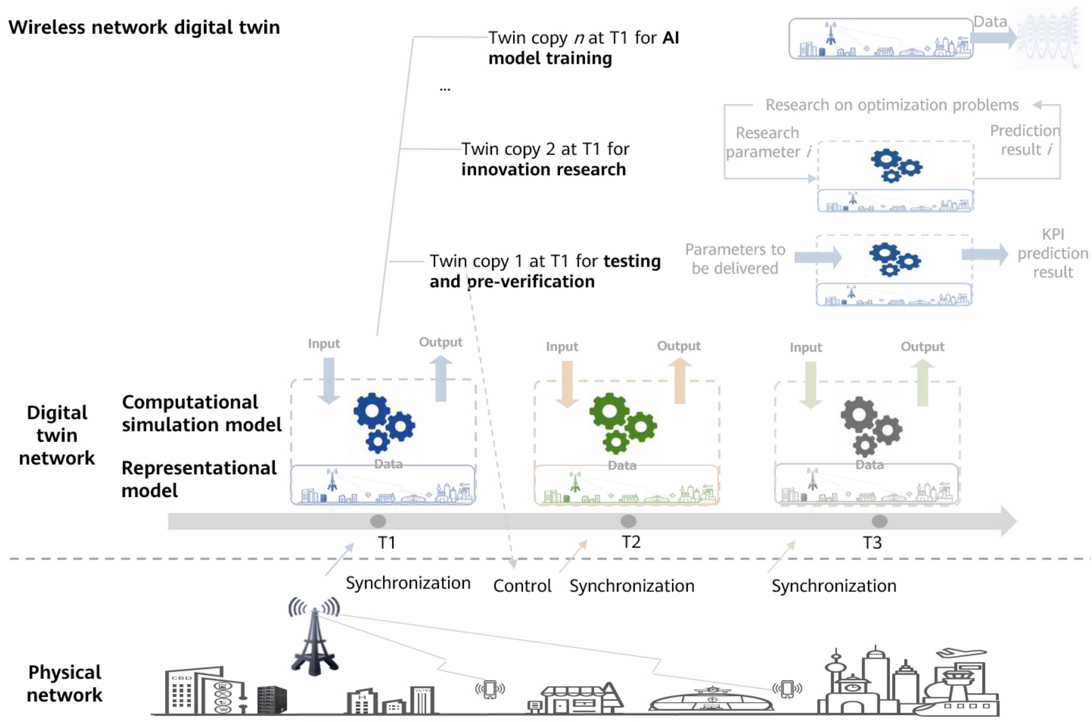
Virtual mesh models for wireless networks include coverage meshes and spectral efficiency meshes. Coverage meshes are designed to avoid any performance degradation caused by inter-frequency measurements. Spectral efficiency meshes, on the other hand, help users accurately identify the capabilities of different carriers. This information is then used as a reference for UEs to select the best carrier to camp on. In this way, the resources of the multi-frequency wireless network can be fully utilized, resulting in an improved user experience.



2.3.2 Digital Twin

Digital Twin is a critical research direction for 5G-A and 6G. As defined by the Digital Twin Consortium, a digital twin is a virtual representation of real-world entities and processes synchronized at a specified frequency and fidelity. Digital twins use historical and real-time data to represent the past and present and simulate predicted futures. Motivated by outcomes, tailored to use cases, driven by integration, built on data, and guided by domain knowledge, digital twins accelerate understanding of the whole system and lead to informed decisions. The Digital Twin Consortium has categorized digital models into two types: representational models and computational simulation models. A representational model consists of structured information and typically represents the state of an entity or process, while a computational simulation model is the executable model of a process and consists of input and output data and algorithms that represent the model.

In the case of the wireless network digital twin, there are two comparable models. The first model, representational model, as shown in the figure below, represents the network in a specific state and is synchronized with the physical network to reflect the entire wireless network in real time or near real time. The second model, computational simulation model, simulates the operating mechanism of the wireless network. Taking into account the status at a particular moment, the model predicts the changes of the wireless network based on the input status data and control actions, and outputs the prediction results after the changes.



2.3.2.1 Representational Model

The wireless network digital twin representational model is a digital representation of the wireless network. Traditional O&M includes data such as configuration, KPI statistics, fault alarms, inventory logs, fine-grained user-level call logs, and operational logs. This data describes specific objects or aspects of the network, but does not provide a complete reflection of the entire wireless network. In order to build an accurate representational model of a wireless network, two technical requirements must be considered:

Entity-oriented modeling: This technique is used to model wireless networks. A network consists of many cells that can be expanded both up and down. O&M data such as configuration, KPI statistics, fault alarms, inventory logs, fine-grained user-level call logs, and operation logs can be attached to entity objects. This creates a digital twin model for each entity object that can be mapped. In addition, entity objects can be associated with each other, and the digital twin model must reflect these associations to allow for elastic expansion of the twin model.

Generate multiple copies of the representational model based on application requirements: The digital twin has the ability to enable different applications, so multiple copies are created based on the data in a particular state on the network. As shown in the previous figure, three copies are created based on the digital twin at T1. These three copies operate independently and do not interfere with each other.

1) Enable pre-verification testing: Before deploying algorithms to the network, a pre-verification test can be performed using the digital twin copy of the current network. This reduces the risks and costs associated with trial and error during live network deployment and improves the efficiency of solution deployment. In addition, the pre-verification capability can be used to pre-evaluate the benefits of optimization policies. This allows users to select the most valuable policy based on the pre-evaluation results.
2) Enable innovation research: Digital twin copies provide a promising basis for algorithm research. For example, one can explore the optimization algorithm for the user-perceived downlink rate in a given area, identify a set of optimal parameter configurations that maximize the rate, and use heuristic optimization or reinforcement learning algorithms to perform hundreds or thousands of interactive iterations with the digital twin environment. The twin environment can provide hourly or daily KPI prediction results that are faster and more reliable than those obtained from live network-based iterations. This enables efficient and cost-effective innovation research. In addition, the digital twin environment can address the challenge of dealing with a large optimization space that includes multiple objectives and methods. As a result, an optimization policy can be generated within a short optimization time.
3) Enable AI model training: The simulation algorithm used in the digital twin generates a large amount of diverse data that can be used as input for AI models. In addition, the AI model can interact with the digital twin environment to receive feedback on the results of each AI algorithm training. In addition to training the model, the AI model can also evaluate the effectiveness of the final AI model.

2.3.2.2 Computational Simulation Model

A computational simulation model of digital twin for wireless networks represents the process. It is expected to comprehensively consider evaluation dimensions such as coverage, capacity, experience, errors, and energy consumption of wireless networks to provide an overall feedback on the networks. The technical requirements are as follows:
1) Design and association of the simulation process: Simulation models are designed separately for the planning, construction, maintenance, optimization, and operation phases of traditional O&M. More research is needed to determine whether to associate and cascade the existing simulation models to form a comprehensive simulation system, or to use one large model to represent each phase.
2) When multiple models are cascaded, an error in a previous level model will

introduce noise into the input of the next level model, causing error propagation. When many models are cascaded, their cumulative error may not increase linearly. In this case, methods such as error compensation and error correction are required to control the error.

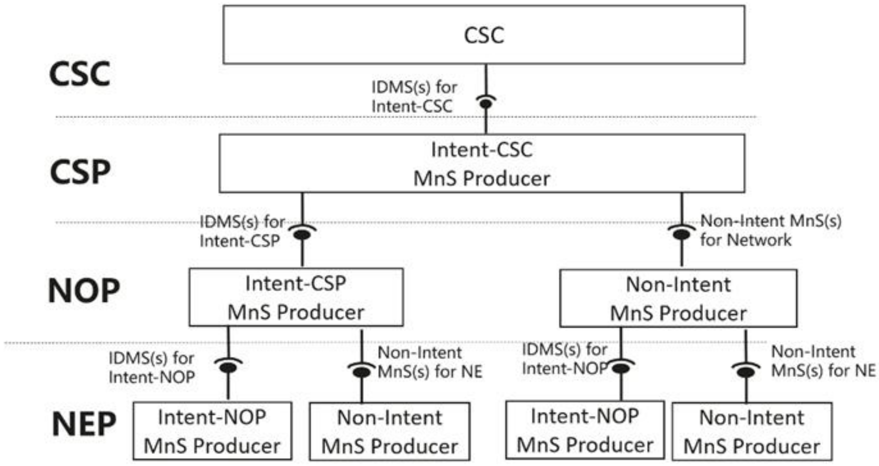
3) When a large model is used, it is still challenging to obtain enough samples and to consider multiple dimensions such as coverage, capacity, experience, errors, and energy consumption.

Comprehensive consideration of computational efficiency and simulation accuracy: When a digital twin network reflects a physical network in virtualized mode to form a twin mirror - just as a mirror reflects the physical world - the total network resources required by the physical and twin networks are doubled. Obviously, this is not an efficient construction mode. Therefore, the actual running programs of the physical network must be processed and abstracted. By using various differentiated data generated by mechanism modeling, AI models are developed, and computational efficiency is greatly improved. However, abstraction may cause a loss of accuracy for some models. It is necessary to strike a balance between computational efficiency and modeling accuracy to eventually transform into a white-box/black-box model driven by both domain knowledge and data.

2.3.3 Intent Openness

In August 2018, 3GPP initiated research on intent-driven management services. Currently, 3GPP Release 18 has started projects on enhanced intent management. As defined in the 3GPP specifications, an intent is a set of expectations including requirements, goals, conditions and constraints given to a 3GPP system, without specifying how to achieve them. Intent openness hides the details of the network implementation from users, meaning that users can obtain appropriate services only by expressing the desired network state in a simple way.

The intent producer automatically extracts key intent information from user input, translates it into appropriate network policy languages, manages conflicts, and delivers the policies to the network for automatic execution. Users can view the intent report provided by the intent producer to obtain the execution status, execution result, and suggested changes of the intent on the network, and intervene in a timely manner to achieve the expected goal. To achieve intent openness, key technologies, such as intent translation, intent evaluation, and intent conflict management must be provided.



2.3.3.1 Intent Translation

Intent translation translates user intents into executable policies that networks follow. 3GPP defines a hierarchical model for intent translation: Intent-communication service customer (Intent-CSC) management service (MnS) Producer receives an intent and translates it into an Intent-communication service provider (Intent-CSP) or a network requirement, and then Intent-CSP MnS(s) or Non-Intent MnS(s) can be used to fulfill the Intent-CSC. The Intent-CSP MnS Producer receives the Intent and translates it into a new Intent for the network operator (NOP) or NE. Intent-NOP MnS Producer receives the expressed intent and translates it into detailed NE requirements. This process converts an abstract intent into detailed NE configuration requirements.

Technically, intent translation involves two steps:

1) Intent recognition: User intents are typically expressed in concise, user-friendly, natural language. The most important step is to accurately extract complete key intent information from user input. Traditional natural language processing (NLP) methods face the challenge of accurately recognizing users' key intent information. This results in intent information loss. Therefore, intent openness often requires users' intent input to be expressed in formatted natural language to facilitate intent recognition. However, with the introduction of large models, their information mining capabilities will be used to recognize important intent information from user input. This will be an important research direction for intent recognition. At present, intent recognition capabilities have been applied to human-machine interaction in the wireless field, simplifying O&M operations and quickly obtaining accurate feedback information.
2) Intent-policy translation: Intents can be satisfied by applying different solutions. Intent-policy translation is an essential step for translating

ambiguous intent information into deterministic policy information. For this reason, the intent producer must have a clear understanding of all solutions to intents supported by the system. In addition, because intent translation technologies are closely related to domains, it is extremely difficult to abstract a common intent translation model for multiple domains. Currently, intent translation is mainly implemented in a single domain (such as energy conservation) based on expert knowledge bases and knowledge graphs. However, due to the massive parameters involved in wireless networks, only certain intents can be translated during a specific implementation process. Large models trained on a large amount of expert knowledge in wireless networks provide a new approach to intent-policy translation.

With the help of large models in the wireless domain, user intents can be accurately recognized and analyzed, and user requirements can be accurately decomposed. Based on the intent translation capabilities, capabilities in each domain can be orchestrated and combined to generate policies that match the intents.

2.3.3.2 Intent Evaluation

Intent evaluation consists of two phases, intent pre-evaluation and intent post-evaluation. Intent pre-evaluation is to estimate the effect of intent execution by employing simulation technologies such as digital twin prior to intent execution, and offer intent delivery suggestions to users based on the effect. Therefore, intent pre-evaluation largely depends on how mature the simulation technologies are. Intent post-evaluation means evaluating the actual fulfillment of intents after they are delivered to networks for execution, and notifying users with an intent report. Intent evaluation herein mostly refers to the latter phase, intention post-evaluation. In the future, intent pre-evaluation will also be an important research direction.

2.3.3.3 Intent Conflict Management

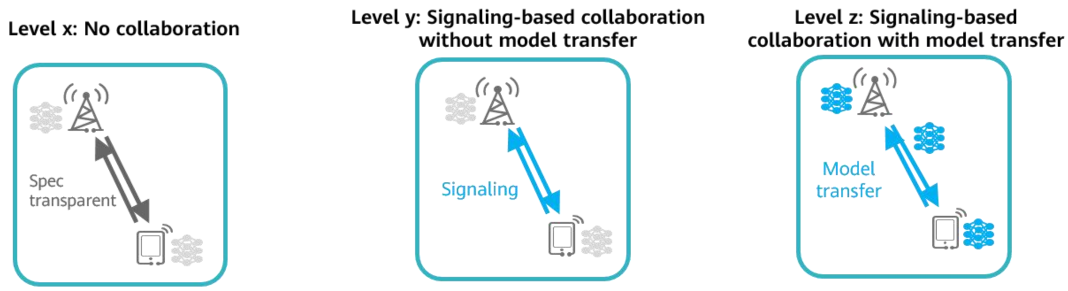
During intent execution, conflicts may arise between multiple intents, between separate intent goals within a single intent, or between new and exiting intents. As a result, there is a need to manage the conflicts. According to the 3GPP specifications, intent conflicts can be divided into explicit and implicit conflicts. Explicit conflicts can be identified using intent models, whereas implicit conflicts need to be identified by intent pre-evaluation and resolved by design optimization.

Currently, intent openness is focused on a single domain. An all-domain intent-driven framework is expected to be established to provide a unified intent entry for all domains and extract common intent translation and intent conflict management capabilities. Until then, users can manage networks from a broader perspective without worrying about the domains of the intents. Intents are automatically decomposed into the appropriate subdomains for collaborative intent execution.

2.3.4 Intelligent Air Interface

An existing wireless communication system is based on a theoretical framework of wireless communication that has been developed over decades. Theoretical modeling is performed on elements in each phase of wireless communication in a mathematical abstraction manner, and an efficient solution is performed within an acceptable error range. This method has been successfully applied to the last five generations of communication systems. However, this modeling analysis method also has some problems. For example, abstract theoretical modeling cannot fully reflect the complex actual radio environment, and the complexity of finding the optimal solution is often very high.

To obtain more accurate information about the radio environment, the network and the terminal may need to cooperate with each other to perform operations such as model training and model updating. Currently, there are three widely accepted levels of collaboration: no collaboration, collaboration with only signaling interaction but no model transmission, and collaboration with model transmission based on signaling interaction.



In Release 18, three key cases of AI-enabled wireless communication are used to guide the industry to think about the value and feasibility of RAN intelligence.

2.3.4.1 AI-Enabled CSI Enhancement

Accurate CSI measurement is critical for downlink data transmission. However, UEs consume a large amount of air interface overhead to provide feedback on measured CSI. The traditional method uses mathematical modeling and CSI sparsity in the angular delay domain to feed back only the CSI compressed in the transform domain. However, this method partially degrades performance. In Release 18, AI or machine learning (ML) is applied to find an optimal mapping rule between the original CSI and the transform domain. The rule can be used to construct an encoder and decoder for CSI compression and restoration, respectively, to make CSI feedback more efficient and greatly enhance the CSI-based edge user experience with the same CSI feedback data.

2.3.4.2 AI-Enabled Beam Management

Millimeter-wave (mmWave) communications, which are characterized by high frequency bands and narrow beams, require beam management to determine an optimal transmit/receive beam. However, the traditional beam management method suffers from problems such as high sweep overhead and difficult beam alignment in high mobility scenarios. In Release 18, AI-based time-domain and space-domain beam management is used to achieve a trade-off between beam sweep overhead and beam alignment accuracy. Spatial-domain beam management follows the idea of super-resolution and predicts the optimal beams among all candidate beams by transmitting only a small number of beams. Time-domain beam management uses historical beam measurement information to predict the optimal beams at a future time. This reduces beam scanning times in mobility scenarios and improves data transmission performance. With the help of AI, fewer resources can be consumed to provide a better user experience.

2.3.4.3 AI-Enabled Positioning

Traditional positioning methods require at least three LOS channel paths to perform triangulation positioning, the performance of which depends on the identification accuracy of the NLOS/LOS channel paths. There is a huge loss of performance in scenarios where NLOS channels are dominant. In Release 18, AI is used to learn radio channel characteristics. When a base station is equipped with only a few antennas, AI will enable high accuracy in identifying NLOS/LOS channel paths in mild NLOS scenarios, providing decimeter-level positioning. In severe NLOS scenarios, AI-based channel fingerprint feature learning can also help achieve decimeter-level positioning.

AI is also characterized by efficient extraction of data features and learning of mappings between data features and optimal decisions. Therefore, networks with native intelligence can be used to sense a wireless environment, extract prior information in channels, and obtain accurate channel information with very few measurements. In addition to radio channel sensing, AI can also be used to sense information such as terminal tracks, postures, and locations, to extract common features from multi-mode information, and to develop more use cases for communication applications such as zero-scan beam management, high-mobility channel prediction, and pilotless channel estimation.

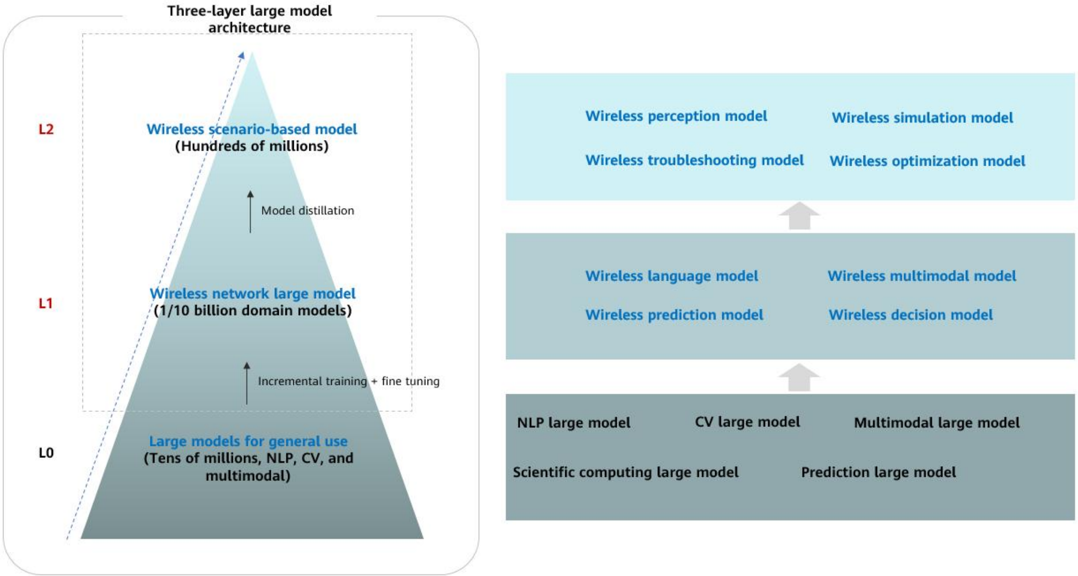
2.4 Potential Technologies for RAN Intelligence - Large Models

Generative Pre-trained Transformer (GPT) large models developed by Artificial General Intelligence (AGI) have opened up new frontiers in Natural Language Processing (NLP). These models can perform various NLP tasks such as intelligent question-answering, text/image/video generation, and summarization/translation. A large language model (LLM) is a type of algorithm that uses deep learning techniques and massive datasets to understand and automatically generate human language. Unlike any conventional natural language model, LLMs can handle multiple NLP tasks. Large models are characterized by pre-training and fine-tuning. They are pre-trained on large textual datasets so that they can perform general functions. To adapt to a wide range of downstream applications, large models are also fine-tuned using a small amount of data.

In addition to the NLP large model, there are also CV and multimodal large models in the industry. Similar to LLMs, CV large models need to understand images so that they can detect, segment, and recognize all objects in the images. CV large models can be used for multiple purposes, including intelligent driving, image recognition, and security protection (facial recognition). Multimodal large models, as a new perceptual and cognitive frontier, can integrate multiple modalities, such as speech, image, and audio. They enable machines to simulate human-to-human interaction by processing and analyzing data from multiple modalities in the environment. In addition, multimodal models can combine information from multiple modalities into a single model and enable the generation of one modality from another.

The introduction of large model technology to the industry provides new approaches to the development of autonomous networks. The large amount of expertise and massive network data in wireless networks is used to train and generate large models suitable for wireless networks. The wireless network-specific models are expected to upgrade the autonomous level of wireless networks to L4, reaching the intent-adaptive network-level intelligence. In addition, the large models can also provide the following substantial benefits: intent perceiving, simplifying man-machine interaction; intelligent assistant, supporting O&M efficiency improvement; digital expert, achieving self-closed-loop in mainstream scenarios; fault prediction and prevention, moving toward "zero-fault" networks.

The large model architecture in the wireless domain consists of the following three layers.



1) In addition to the lowest layer (layer 0), the large model architecture for wireless networks is also defined with two other layers (layer 1 and layer 2). The wireless network large models are understood as level 1. Most basic models dedicated for wireless networks are located at layer 1, such as the LLM, the multimodal large model, the prediction model, and the decision-making model. Moving up one layer, we are talking about the wireless large models for specific industry scenarios, which are developed based on the layer 1 models. Layer 2 wireless network models, including the perception large model, the simulation large model, the troubleshooting large model, and the optimization large model, are designed to perform specific tasks.
2) The wireless network language large model (layer 1), a model driven by and trained on natural language data, is generated by incrementally training and fine-tuning the NLP model. To train the NLP model, large textual datasets consisting of hundreds of billions of corpuses are required. The model is designed to be equivalent to GPTx, the term used to refer to large language models derived from OpenAI's series of GPT releases. Hundreds of millions of corpuses, tens of thousands of question-answer pairs, and chain-of-reasoning data are required to build a wireless language large model. Massive wireless data is fed into the model to help it understand all the knowledge in wireless networks. Furthermore, the model can be used to generate knowledge based on logs, stimulate a variety of faults with the existing knowledge, and provide one-stop O&M capabilities in optimization and deployment scenarios.
3) Intent-driven O&M is a key feature of ADN at L4, serving as a bridge between customers and wireless network O&M. As a key technology of the knowledge-driven wireless network language large model, intent-driven O&M, unlike the traditional O&M mode, can understand customers' intent expressions and translate them into intent interfaces, invoke existing APIs

or functions to orchestrate and complete tasks, and provide summary or analysis reports, thereby improving O&M efficiency.

In conclusion, the emergence of large model technology in the industry provides new approaches to the development of autonomous networks. The large amount of expertise and massive network data in wireless networks is used to train and generate large models suitable for wireless networks. The wireless network-specific models are expected to upgrade the autonomous level of wireless networks to L4, reaching the intent-adaptive network-level intelligence. The layer 1 basic models for wireless networks, such as the language large model, multimodal large model, prediction model, and decision-making model, have made significant contributions to the development of autonomous networks. These models aim to improve the performance of wireless communication systems by efficiently extracting data features and learning the mapping between data features and optimal decisions.

2.5 Data Requirements for Wireless NEs (OMC and Base Stations)

1) Data Type

Intelligent applications are classified into O&M applications and NE applications. O&M applications obtain their data from the OMC and external systems, while NE applications obtain their data from UEs and base stations.

Application Type	Data Type	Data Source	Data Usage	Granularity
O&M applications	Traffic statistics	OMC	OMC	Within hours
	Configuration data	OMC	OMC	-
	Alarms	OMC	OMC	-
	CHR/MR	OMC	OMC	Within seconds
	Engineering parameters	External data	OMC	-
	E-maps	External data	OMC	-
NE applications	Air interface channel measurement data	UE	Base station	TTI
	Traffic volume/Number of UEs	Base station	Base station	Within seconds

2) Air Interface Standardization

To enable air interface intelligence during the evolution to 5G-A, the signaling protocols between base stations and UEs need to be defined in 3GPP specifications. Specifically, the air interface signaling process at Level x/Level y/ Level z needs to be designed in RAN 1/2/3.

3) Interface Definition and Openness

To improve E2E UE experience, external wireless data must participate in the joint decision-making of the wireless system. It is recommended that external data collection channels and data content within the scope of operator authorization be further defined in 3GPP specifications.

3 USE CASES OF RAN INTELLIGENCE

3.1 Intelligent Multi-Band Coordination

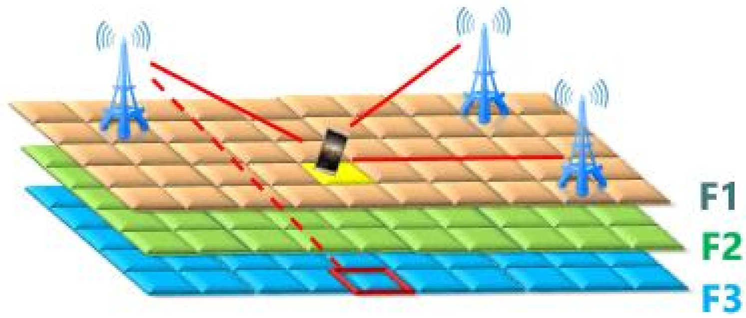
Application Scenarios

When networks operate in multiple frequency bands with both Frequency Division Duplexing (FDD) + Time Division Duplexing (TDD), TDD + FDD, or FDD + FDD transmission schemes, UEs are typically handed over between different frequencies to obtain an optimal frequency for camping. Inter-frequency handovers require inter-frequency measurements to obtain signal strength data for reference. However, inter-frequency measurements lead to reduced UE throughput, especially for Cell Edge Users (CEUs). In addition, due to the latency of inter-frequency measurements, an optimal carrier cannot be selected for UEs in a timely manner. To address this issue, AI technology is used to build a coverage grid model whose prediction result is used to replace the inter-frequency measurement result.

Based on whether a neighboring frequency exists during a handover, there is inter-frequency coverage or such coverage does not exist.. When the inter-frequency coverage does not exist, the UE cannot detect inter-frequency neighbor cell signals by the traditional method of providing inter-frequency measurements. As a result, the UE is expected to continuously detect the signal, which severely affects the UE throughput. When there is inter-frequency coverage, a measurement can successfully obtain required inter-frequency information with the conventional measurement mode. A single inter-frequency measurement has limited impact on the UE throughput, but the latency of the handover procedure is large.

Key Technologies

Based on the multi-dimensional measurements of radio signals, UEs with the same characteristics are categorized into a group. Base stations consider UEs with the same measurement results to be in the same grid, referred to as a virtual grid-based network. For example, if the measurement result on a frequency reported by UE 1 and UE2 are [(Cell 1, RSRP 1), (Cell 2, RSRP 2), ...], the two UEs are considered to be in the same virtual grid. Based on the intra-frequency and inter-frequency MR data of historical UEs in a cell, base stations use supervised machine learning technology to discover the association from intra-frequency signals to inter-frequency signals. In this way, the base stations can build coverage grid models. After base stations obtain the virtual grid information about a UE on a frequency, the base stations query the coverage grid model for a neighboring frequency to quickly predict the RSRP of the UE on that neighboring frequency. The prediction result can be used for inter-frequency handovers. In this way, inter-frequency measurements are not required.



Capability Requirements
Coverage grids need to be built and inter-frequency measurement information can be predicted based on intra-frequency information.

Data Requirements
Intra-frequency and inter-frequency measurement information reported by a certain number of UEs is required.

Expected Gains

When there is no inter-frequency coverage, the UE throughput increases by about 10% using the coverage grid for prediction compared to the traditional inter-frequency measurement method. When inter-frequency coverage exists, the latency of the coverage grid-based handover procedure is about 300 ms shorter than that of the traditional inter-frequency measurement procedure.

3.2 Designated Application Experience Guarantee

Application Scenarios

As the impact of new technologies, new models, and new forms of business on traditional industries continues to intensify, digital transformation has become a consensus among global enterprises. As the scale of 5G networks continues to grow, the growth of data services also poses higher requirements for network performance. By providing fine-grained protection for services, it can meet the differentiated experience requirements of different services, thereby further promoting 5G user development, bringing about the improvement of terminal penetration rate and the expansion of business scenarios, truly releasing the potential of 5G networks and enhancing the value of 5G networks.

Key Technologies

- Fine-grained recognition. By embedding a deep recognition module in the base station side, it can accurately identify services, and the specific level of recognition can be determined by customer needs.
- Precise guarantee. By prioritizing and guaranteeing the rate of the identified services, it can achieve differentiated guarantee.
- Perception quantification. By establishing an xEMI model to quantify and score different services, it can achieve accurate service perception.
- Problem optimization. It can quickly delimit and locate poor-quality users, and achieve cell/CDR/user/grid-level multi-dimensional statistics, combined with MR/CDT and other data correlation analysis, locate the wireless root cause of poor xEMI; and it can perform detailed analysis and location of service perception abnormal events, and provide optimization suggestions.

Capability and Data Requirements

This technology application requires adding an operation and maintenance server NGI, upgrading the wireless base station software version, and has no impact on other network elements such as terminals and core networks.

This technology application can be opened to operators through atomic capabilities, and improve user perception and network traffic. It can also be opened through wireless APP-level perception indicators, and help operators build a wireless APP-level perception experience evaluation data system.

3.3 Intelligent User Orchestration

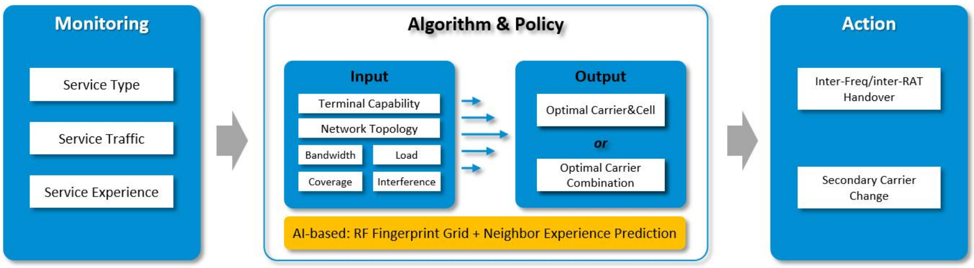
Application Scenarios

In the 5G era, mobile communication networks have developed to an extremely complex stage. The network form is flexible and diverse, with more frequency band resources (4G/5G, low frequency/high frequency) and more network service modes (LTE/SA/NSA/CA), among which NSA also includes various dual connection modes; the business needs are rich and varied, the traditional 2C business continues to evolve, and the promising 2B business is widely used in various industries; the terminal capability gap widens, from the stock 4G terminals to the rapidly iterating 5G terminals, as well as various industrial terminals and IoT modules, as the 3GPP protocol version continues to evolve, the network needs to adapt to various differentiated terminal capabilities; the wireless environment is complex and changeable, covering, interference, load and other issues are all intertwined, and the network optimization work is intricate.

In such a complex situation, how to flexibly adapt to various network forms, business needs, terminal capabilities and wireless environments, select the optimal frequency or frequency combination for users, as well as the optimal service mode, while taking into account both the individual experience of users and the overall performance of the network? This poses a huge challenge for 5G network operators and base station equipment manufacturers.

Key Technologies

Intelligent user orchestration is driven by user perception improvement as the core driving force, comprehensively considering terminal capabilities and business needs, terminal's spatial-temporal location and the service capabilities that the network can provide at that location, and combining with wireless fingerprint grid knowledge base atomic capabilities, to obtain the optimal solution from the given frequency band & cell set and network service mode set, that is, to orchestrate the optimal network service mode, optimal frequency or frequency combination, and optimal service cell for UE



Functional Framework of Intelligent User Orchestration

Capability and Data Requirements

This technology application does not require network transformation, only wireless base station and network management software upgrade, and has no impact on other network elements such as terminals and core networks. It opens the "intelligent user orchestration" atomic capability to operators, and the upper layer network management of operators can call the wireless single-domain atomic capability through open interfaces such as OpenAPI, to achieve user experience improvement goals.

3.4 Intent-driven Service Experience Guarantee

Application Scenarios

The communication industry, as a basic industry for national economy and people's livelihood, shoulders the mission of safeguarding the national construction and the normal operation of thousands of industries, and is the cornerstone of the entire social development. And in daily life, communication guarantee is the most common scenario, from important national meetings and earthquake relief, to people's entertainment life and sports events, there are often major events that require operators to ensure the normal development of communication activities, avoid communication failures caused by emergencies, and thus cause huge negative impacts on the entire major event.

Traditional communication service guarantee requires high technical threshold for guarantee engineers, and it is difficult to respond to business changes and emergencies in time; and it cannot implement guarantee for service-level experience, and cannot flexibly adjust guarantee strategies with the changes of network conditions. With the widespread introduction and application of intelligent technology in the 5G era, the demand for intelligence, simplification and de-risking of communication guarantee is also increasingly urgent. Under such a background, the time is ripe for intent-driven integration into various mainstream business scenarios in the communication field.

Key Technologies

Intent-driven service experience guarantee, through natural language input of service experience goals, the system automatically translates the intent rules to guarantee strategy mapping, real-time monitors the guarantee effect and adaptively adjusts the optimal guarantee strategy, and improves user service experience.

- Intent input. By abstracting the intent input model, forming a four-tuple expression of region, APP, time, guarantee level, and based on the DSL modeling method to improve and build a complete set of descriptions for the intent domain, the intent expression is changed from the traditional imperative to declarative way, completely changing the way of human-machine and machine-machine interface transmission.
- Intent translation. It includes accurate translation of natural language to intent rules, accurate translation of intent rules to guarantee strategies, and automatic recognition of services. Based on natural language models, such as Bert, it completes the automatic recognition library of communication and geographic domains, and realizes end-to-end translation from specific operation and maintenance to business scenarios.
- Conflict detection. It automatically detects and avoids conflicts between multiple intent tasks, including detecting whether the execution of perceived service guarantee intents is feasible, whether there are conflicts with other services, etc., predicting and managing the reachability and conflicts of intents, and providing modification suggestions.
- Intent execution. After the intent task is translated and conflict detected, it distributes the intent goal to the base station for execution. The base

station provides different levels of guarantee capabilities, realizes differentiated service recognition and quality guarantee, including real-time evaluation of perceived MoS for service CDRs, thereby estimating the quality of service perception experience. If it is lower than the target, it performs closed-loop optimization for the services that need to be guaranteed according to the difference.

- Intent guarding. During the execution of the intent task, it provides full life cycle performance monitoring for the guaranteed services, updates the root cause recognition of service recognition and congestion in real time, evaluates the execution effect of the intent, and adjusts the guarantee strategy in a closed loop according to the effect.

Capability and Data Requirements

This technology application does not require network transformation, only wireless base station and network management software upgrade, and has no impact on other network elements such as terminals and core networks.

It opens the intent-driven service experience guarantee atomic capability to operators, and the upper layer network management of operators can call the wireless single-domain atomic capability through open interfaces such as OpenAPI, and through clear intent goals, make the system achieve better guarantee performance, and the upper layer network management can also easily judge the intent achievement effect through intent feedback.

3.5 Intelligent Macro-Micro Collaboration

Application Scenarios

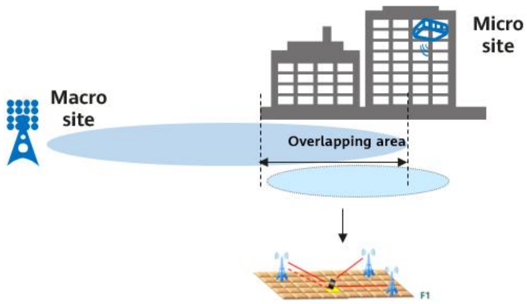
When a UE moves from an indoor cell to an overlapping area between outdoor and indoor cells served by macro and micro base stations operating on the same frequency, the UE may not get the optimal downlink experience after selecting the cell with the largest RSRP to camp on. This is because the macro and micro base stations have very different antenna gains. This feature allows UEs to access cells with good spectral efficiency and UE experience, significantly improving UE experience in overlap areas.

When a base station selects carriers for a UE, the base station determines whether to change the carriers for the UE mainly based on the RSRP reported by the UE. In this case, the selected carriers may not be able to provide an optimal UE experience. The AI-based spectral efficiency grid model can predict the spectral efficiency of individual UEs on different carriers and can therefore be used to help select carriers that can provide optimal UE experience.

Key Technologies

Based on the intra-frequency MR data, UE capability, and spectral efficiency of

previous UEs in a cell, base stations use supervised machine learning technology to discover the association between the data. In this way, the base stations can build spectral efficiency grid-based network models.
After obtaining the virtual grid information of UEs in the overlapping area between macro and micro base stations, base stations query the downlink spectral efficiency models of the macro base station and micro base stations to predict and compare the spectral efficiency when UEs access the macro and micro base stations. Then, the optimal site is selected for the UEs to camp on.



Capability Requirements

The spectral efficiency grid model can be constructed, and the spectral efficiency can be predicted based on information such as intra-frequency measurement and UE capability.

Data Requirements

Intra-frequency measurement information, UE capability information, and spectral efficiency information reported by a certain number of UEs are required.

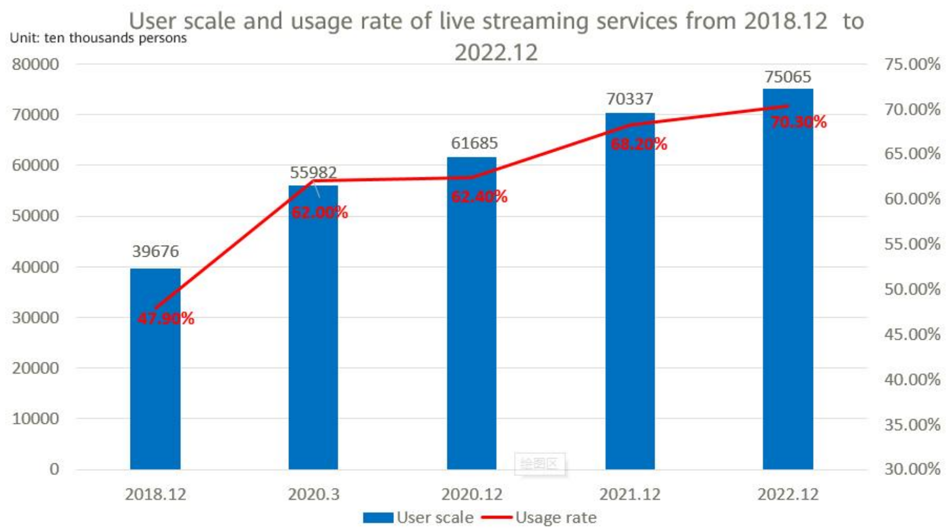
Expected Gains

In typical traffic scenarios, if there are enough UEs in the overlap area between intra-frequency macro and micro base stations and the rate difference between macro and micro base stations is large, the spectral efficiency grid can be used to select a site with the best experience for UEs in the overlap area to camp on. This increases the UE's perceived rate in the overlap area by 5% to 10%.

3.6 Live Streaming Service Experience Assurance

Application Scenarios

The emergence of live streaming services is expected to increase the number of live streaming UEs, which will increase traffic by N times. The quality of live streaming services largely depends on the rate and resolution improvement and latency reduction of the mobile network. If poor UE experience, such as frame freezing, occurs, tens of thousands of revenue can be lost. By the end of 2022, the number of online live streaming UEs will reach 751 million, accounting for 70.3% of netizens.



Source: Statistical Report on Internet Development in China

To ensure the live streaming service rate, the service experience data sets are used to train a model that aims to visualize the poor QoE perception and improve the accuracy of its root cause analysis. In addition, there are other SLA service requirements on the network. For example, cloud games are sensitive to latency. As a result, the intelligent multi-objective optimization capability is integrated for proper resource allocation to meet the SLA requirements of different services.

Key Technologies

Necessary features can be extracted from parameters such as power control, bit rate, and frame rate of a base station to model the rate of a virtual network. This model helps predict the UE service rate. By taking into account factors such as different service packet sizes, the impact of UE power on spectral efficiency, neighboring cell loads, and the impact of UE distribution on interference, it is possible to accurately predict the rate change of live streaming services. In this way, key factors that degrade the UE experience can be adjusted and optimized in advance. 5QI-level multi-objective parameter optimization is performed based on the prediction result of the service rate of a virtual grid-based network. This capability not only satisfies the UE rate requirements in the optimization cell and the target cell, but also ensures the experience of background UEs in the cell.

Capability Requirements

Dedicated 5QIs are configured for live streaming UEs. The OMC system can construct a model to predict the rate of a grid-based network and quickly adjust UE group-level parameters based on poor-QoE analysis.

Data Requirements

Base stations subscribe to cell-level and 5QI-level traffic statistics and CHR data.

Expected Gains

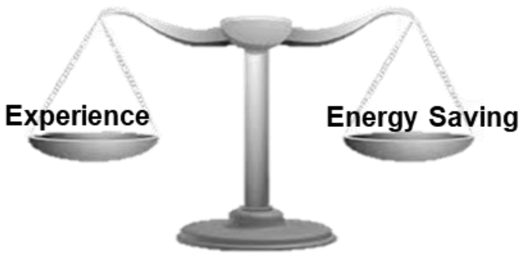
The UE-perceived rate of live streaming services is expected to increase by 5% to 10%.

3.7 Energy Saving While Maintaining a Stable Network Performance

Application Scenarios

Due to the increased network size, network complexity and workload, 5G networks today have a much higher OPEX and consume more power, with an annual power bill of over CNY10 billion. To address this situation, energy conservation remains a top priority for the entire industry. To achieve more energy savings, many radical practices, even cell deactivation and power failure, are adopted. Obviously, this solution is not optimal as it results in a worse UE experience. It is essential to achieve a balance between power consumption and bit efficiency, which will be a major challenge in the coming days. Traditional energy savings solutions are designed to help operators achieve the greatest energy savings while maintaining KPIs and stable network performance.

For one thing, the higher the workload and traffic, the worse the UE experience on the live network. During off-peak hours with lower workloads and traffic, the UE experience is much better or even x times better than during peak hours. Compared to energy-saving solutions during off-peak hours, peak-hour solutions tend to be more conservative. It is important to keep the UE experience as close to the same as possible, regardless of whether it is peak or off-peak. On the other hand, whether energy-saving solutions are aggressive or conservative results in completely different UE experiences. Operators today are working to find an optimal energy savings solution to reduce power consumption without compromising the UE experience, aiming for a win-win outcome.



User experience
imperceptible

Key Technologies

Necessary features can be extracted from the workload, UE distribution, coverage, interference, and energy consumption of base stations to train a model that helps accurately predict energy savings and performance simulation. This model can also be used to accurately predict energy consumption and rate changes as different energy saving policies and performance parameters are developed. When parameters are combined on a large scale under the guidance of multi-objective coordinated optimization, operators strive to optimize network performance while reducing energy consumption and extending coverage.

Capability Requirements

Operators are able to train a simulation model for predicting wireless network performance and energy consumption as well as producing energy savings, enlarging coverage, and improving UE experience.

Data Requirements

Base stations subscribe to cell-level traffic statistics and CHR data.

Expected Gains

Energy consumption is expected to be reduced by 5% to 10% at the network level.

3.8 Network Fault Prevention and Prediction

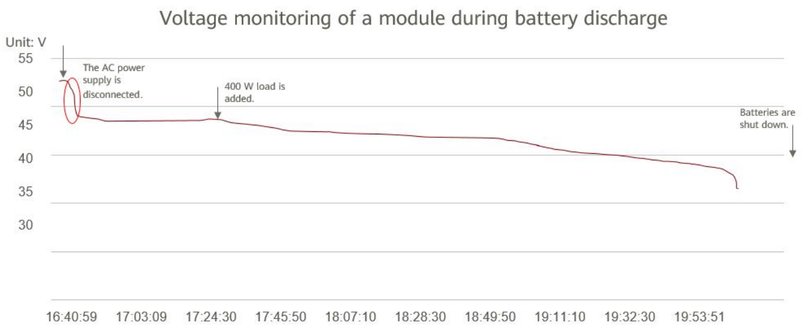
Application Scenarios

In the 5G era, operators are primarily focused on keeping services online at all times. Against this backdrop, network O&M assurance is expected to be a major issue. To bringing us one step closer to "zero fault and always online" wireless networks, the industry agrees that it is important to replace passive responses with proactive prediction and prevention.

A large number of alarms are reported on the live network. A work order is submitted for each fault, consuming a large amount of manpower. With intelligent fault prediction and prevention, more than 80% of potential risks can be detected in advance, enabling proactive O&M and moving towards a zero-fault wireless network.

Key Technologies

The prediction and prevention solution collects the multi-dimensional status data of each device on the NEs in real time and reports the data to OMC. Based on long-term historical statistics, the OMC system uses intelligent modeling algorithms and generates a baseline failure prediction model through training. Together with the real-time data of a single site, the system performs inference to achieve the optimal predictive effect and identify potential failures in advance. The backup power duration prediction solution is used as an example. At each stage of battery discharge, the voltage change of the radio frequency (RF) module is detected to identify the power supply status. When the main power is switched to the backup power, the intelligent prediction and prevention function can be used to model the voltages at the point of exhaustion of various backup power supplies. The goal is to accurately predict the duration of the backup power supply and identify potential outages in advance. The predictive capability can also be used to identify potential optical path risks, high temperature risks, and transmission risks. Long-term and short-term measurement data from OMCs and NEs can be used to predict and determine the subhealth status of optical components, paths, and devices.



Capability Requirements

The feature needs to run on the live network for a period of time (at least three days) for online learning and model training.

Data Requirements

OMC can subscribe to 15-minute traffic statistics and report FDRs, configurations, MML commands, and alarms.

Expected Gains

Failures can be predicted in advance, and possible causes and troubleshooting guidance can be provided. Troubleshooting in advance can effectively reduce the risk of failure. Take the example of predicting the duration of backup power. The number of work orders for downtime caused by backup power exhaustion is about 15% to 30%. Predicting the duration of backup power can greatly reduce the number of work orders caused by this reason.

3.9 Network intelligent abnormal status monitoring

Application Scenarios

5G base station is a complex system composed of a series of hardware equipment, including base station antennas, radio equipment, transmission equipment, etc. Compared with traditional 4G base stations, 5G base stations have larger cell bandwidth, higher power and more complex system structure, requiring greater research and development investment and more maintenance work. In actual use, there will be problems such as processor death, core death, abnormally high interference, zero traffic, steep drop in rate, and no access service, which will affect the quality of network service. At present, such problems can only rely on the manual operation and maintenance mode. In order to timely discover and solve problems as soon as possible, operation and maintenance engineers are often required to monitor and analyze network KPIs in real time with a large workload, resulting in long problem analysis cycle, poor timeliness and high cost.

Key Technologies

By deploying the network intelligent abnormal status monitoring module in the intelligent operation and maintenance system, the hidden fault threshold monitoring function is provided. It is based on soft probe technology and integrated DPI, through the collection of original counter, combined with expert experience, operation and maintenance experience and AI algorithm, all 4G, 5G network involved in equipment status indicators and business performance indicators can be carried out 24 hours a day unattended intelligent monitoring. It adopts task management mode, and users can customize monitoring tasks, monitoring parameters, and monitoring cycles. After activation, the system runs according to the defined monitoring granularity, and automatically identifies site equipment or cell faults. For VIP guarantee sites, it also supports advantage rendering to identify the priority of problem sites. The monitoring results can be dynamically displayed on the interface, and reports can be exported in real time or periodically. In addition, reports can be reported to the ftp server, and third-party tools can be used to send messages such as wechat messages. The monitoring module also provides a flexible, readable, easy-to-understand condition editor and a hidden fault warning rule base,

which can adapt to various complex environments and complex business logic judgment in the field.
By analyzing the results, users can quickly find and locate problems.

The network intelligent abnormal status monitoring module provides users with a flexible, efficient and convenient intelligent monitoring tool, which can accurately identify the hidden fault of the base station, timely and accurate early warning, liberate operation and maintenance manpower, and improve operation and maintenance efficiency. Provide efficient monitoring, the whole network can complete monitoring and early warning within 10 to 15 minutes, give full play to the network management platform data concentration, flexible calculation, friendly interface characteristics, timely and accurate warning base station and community KPI abnormalities, base station hanging and other hidden faults. To solve the problems such as large scale analysis of the field manual analysis cell, time-consuming analysis, too many counters can not be fully monitored, and complex multi-indicator correlation scenes.

When the module finds abnormal indicators or unexpected fluctuations, it can issue timely warnings, and provide detailed and effective early warning information for the reference of operation and network optimization engineers, so that network problems can be solved in a timely manner without the user feeling, and improve the service quality of network equipment. In addition, it can effectively reduce operation and maintenance costs and save human resources while improving operation and maintenance efficiency.