GTI AUTONOMOUS NETWORK V3.0

Towards L4+ Highly Autonomy

Table of Contents

1. Introduction

1.1 Background

In 2020, GTI has established the intelligent network program (renamed in 2020, as "autonomous network"), to provide a reference approach for operators' digital transformation through injecting intelligence deeply into their networks. With large-scale practices in the past 3 years and joint efforts and contributions from service providers, vendors and standard organizations, common census has been widely reached on its vision, target architecture and evaluation system. And more and more CSPs have released their own AN realization approach and strategic goal of reaching L4 autonomy by 2025, including China Mobile[1], China Unicom[2], China telecom[3], AIS[4], Telecom Argentina[4] and MTN[5]. "Enabling self-X capabilities, delivering zero-X experiences." Autonomous networks have not only been an option for operators to realize operation efficiency increase and cost reduction, but a critical pass into committing users with new promises, exploring emerging business opportunities and unleashing potential growth in a broader market.

The 1st [6]release of this whitepaper has drawn a vision for autonomous network, answering the question "what autonomous network should be" and "what problems do we expect it to solve". The use cases shed light on the points where the intelligent transformation of network might begin with. It defines the general evaluation methodology of network autonomous level, reference autonomous network architecture and corresponding requirements on network elements and network management system. The 2nd [7]release went deeper and more concrete with the evaluation method, and further focuses on the 4 major autonomous capabilities, which are autonomous perception, autonomous diagnosis, autonomous prediction and autonomous control. It also highlights some early studies of trust in AN and AN life cycle management. The use cases addressed in these 2 releases are mostly L2/L3 autonomous network solution. This whitepaper will focus on the key technologies and high value use cases enabling the L4 autonomy, with expectations to pool strengths from all parties to solve the challenges and fully unleash the potential of those new trending and solutions to help the industry moving its way forward.

1.2 Overview

The previous work on the baseline of Autonomous networks, including 4 aspects from AN evaluation, architecture to AN elements and management can be found more details and examples in [7].

As the industrial practices entering a whole new phase, this 3rd release will serve as an update on industrial progress, including the latest work in leading CSPs, SDOs and end-to-end solutions from vendors towards L4 advanced autonomous network. It will also analyze the problems and bottlenecks faced by operators in this stage of implementation, and highlight the emerging technologies and breakthroughs that become new engine driving the iterative evolution. Sincere thanks to all the contributors and the supporters from China Mobile, CICT, Ericsson, Huawei, Nokia

and ZTE for their hard work in this white paper, especially for Ms. SONG Xiaojia and Dr.CAO Xi, for their previous efforts in laying a solid foundation for this project. Respectfully, the task leaders and contributors are listed as following.

China Mobile
Lingli DENG, Kaixi LIU, Yan YANG, Chengcheng FENG.
Xi CAO, Xiaojia SONG.
CICT
Xuan ZHANG
Ericsson
Bo GUAN, Meng LAI
Huawei
Li WANG , Yu LAN , Jianqiu XIONG, Jun GAO, Danni CHEN, Ruiyue XV
Nokia
Zhuyan ZHAO, Fangfang GU
ZTE
Yi DING, Yong ZHAN

This is the third version of GTI Autonomous network whitepaper, it will be continuously updated according to the research and development progress.

2. Industry Outlook

Through systematically introduced automation and intelligent technology, autonomous network builds self-configuring, self-healing and self-optimizing network capabilities, and creates zero-waiting, zero-failure and zero-contact customer experience, aiming to reduce costs, increase OAM efficiency, and expand business income.

2.1 General introduction

Through the explorations and practices in the past few years, telecom industry has reached agreement on autonomous network transformation. Related investment and application development are in early stages with an expected future. The latest survey by TM Forum[8] shows that, more than 85% of CSP have started on autonomous network transformation and are currently in the initial stages of L1-L2 autonomy, driven by two major factors: increasing efficiency and reducing costs, improving experience and expanding business. Although the proportion of autonomous network assets investment is generally low, more than 90% of CSPs are in favor of autonomous network vision, and one third of them have made large-scale investment strategies and plans for it. More than half of CSPs expect that AN investment will double in next 3-5 years.

2.2 Best practices of CSP: large-scale deployment towards L4 by 2025

Domestic

China Mobile has been recognized as the world's most advanced autonomous network operator.[8] As the initiator of the AN concept and the pioneer of AN deployment, China Mobile took the lead in putting forward the vision, being the 1st CSP to propose L4 goal, initiated the systematic implementation strategy from network OAM to collaborative NMS and network element evolution, and innovatively put forward the target AN architecture, which was guided by the two business objectives of "business growth and efficiency increase", realized three closed-loop of "resource scheduling, business operation and customer demand" and built the digital intelligent capability at 4 layers of "network element, network, business, services". China Mobile practices the iterative closed-loop capability development method of “industry promotion, top-level design, system development, evaluation and analysis”, and achieved remarkable results. The autonomous level of whole network has been upgraded to L2+ at the end of 2021. China Mobile has widely been recognized in the industry, awarded with various honor including TM Forum Excellence Award 2 in a row, CX Telecom Annual Progress Award and Best SNAI use case of 2021 by CCSA TC610, etc.

China Unicom put forward the zero-X and self-X vision for users and frontline OAM: zero-waiting, zero-failure, zero-contact, zero-risk and self-planning (including base station traffic improvement, site selection/planning efficiency), self-configuration, self-optimization and self-healing. A three-layer architecture of "application, platform and network " is proposed to form a cross-layer and cross-domain collaborative closed loop. China Unicom practices a "trinity" implementation methodology including ANL evaluation, product development and intelligent empowerment. China Unicom’s AN practice is also deployed large-scale on the whole network, identifying and planning the automation/intelligence capability in whole lifecycle of network operation in a top-bottom way, and realizing the improvement from single point capability to process capability, then to whole scenario capability through development of CT/IT products and technologies.

China Telecom takes " Ultra-premium customer experience, ultra-fast telecom services and ultra-intelligent cloud network operations" as its target vision. The company is focusing on all production scenarios of cloud network operations and follows a “cyclic evolution in four dimensions and five steps” methodology: top-level design, level evaluation and shortcoming identification, strategy formulation of AN capability improvement, pilot verification, and AN capability construction. China Telecom is also piloting level evaluation at provincial level, optimizing the evaluation standards, improving their feasibility and integrity, and promoting evaluation and autonomous cloud network operations in the group and all 31 provincial subsidiaries across China.

International

AIS, the biggest CSP in Thailand, released the cognitive telco strategy in 2022, with the vision of " beyond zero-contact operation", and proposed to achieve the L4 autonomous goal by 2025. Based on the dual driving engines of AN and IT intelligence to realize zero-contact operation, referring to TMF's target architecture, AIS's AN practice focuses on two scenarios of "fault management" and "customer complaint" to carry out innovation, and already defined the target of reaching L3 autonomy in 2023 and detailed characteristics in specific scenarios.

MTN, a leading CSP in Africa, recognized and based on TMF' AN vision, released their own "ambition2025" strategy, that is, analysis and decision making based on predictive or active closed-loop management. MTN defines the ANL standard for the whole life cycle of "network planning, deployment, maintenance, optimization and operation", and refers to the iterative cycle methodology, formulates the overall promotion plan "Master Plan" for AN implementation, to build the next generation digital OSS (Operation Support System) platform as a digital engine supporting man-machine collaboration, and support future AI capabilities.

Telecom Argentina defined its vision of optimizing workflow, reducing CAPEX and OPEX, operating more efficiently, NaaS and PaaS, and proposed a five-layer target architecture of infrastructure, infrastructure management, domain orchestration, business orchestration and customer management. Telecom Argentina started the formulation of ANL evaluation standards in 2021, and their autonomy level is rated as L1+. In the future, TA plans to further focus on specific scenarios, and promote innovation and practice according to the AN implementation method of "top-level design, capability evaluation and analysis, capability development and application".

The above practices of various CSPs show that AN deployment is a collaborative evolution of business layer, service layer and resource layer guided by overall strategy. By further defining a set of effectiveness indicator, 3 type of core capabilities, namely, open interfaces, closed-loop orchestration and full stack, can be built on two levels of resource autonomous domain and service autonomous domain in a collaborative way.

2.3 SDO Progress: enhanced collaboration to guide cross-domain standards formulation

With industry practices going broader and deeper, it becomes a common recognition that the key to accelerate AN implementation is enhanced evolution of cross-layer integration API and standards formulation. Standards are the key element to promote the cross-layer collaborative development. The research and promotion of cross-layer API-related standards are particularly worthy of attention, for they can improve AN capability in the autonomous domain of network elements, ensure the automation and intelligent processing of network elements themselves, and open intelligent capabilities for cross-domain collaboration such as parameter adjustment, service provisioning and path optimization on OAM level.
A wide range of standard around AN has been launched in the past 3 years, covering business and technical architecture, ANL, and key technologies/interfaces, together with third-party evaluation and certification programs in the past year to test the network equipment of vendors and CSP's service.

TM forum has so far the most comprehensive AN standards layout covering all domain. The delivered assets include business architecture9, autonomous network level evaluation standard and methodology[10], technical architecture[11] and intent in AN[12]. In 2021, it has also launched a multi-SDO collaboration project. With TMF as the core, the project is expected to guide cross-domain AN standard development with ETSI, CCSA, 3GPP, ITU-T and other organizations. Since 2019, TM Forum has released 4 versions of Autonomous Network Whitepaper, with more than 50 industry partners engaged.
3GPP SA WG5 has launched a series of projects on autonomous network, covering whole lifecycle of network management. Release 17, frozen in 2022 Q2, includes intent driven management service of mobile network, management service assured by communication service, study on enhanced management data analysis[13] and 5G SON[14]. Currently, agreement has been widely reached on the layered and domain-specific network architecture[15-17], key concepts[18], use cases and solutions on level evaluation of autonomous network, definition on autonomous network level[19], autonomous network closed-loop control, concept, application and solutions on intent driven management[20-21].

CCSA, as the most important domestic standardization organization in telecom industry, has formulated the framework of a series of standards for "Intelligent Operation and Management of Information and Communication Networks", covering the architecture[22], use cases, level evaluation[23], intent management[24-25], policy management, closed loop management, knowledge management[26] and other key technologies of autonomous networks. By October 2022, there has been 6 joint meetings of CCSA TC7. 29 AN related technical specifications projects and 10 technical reports projects have been approved, with both general level evaluation standards [27]and domain-specific level evaluation standards[28]. In order to speed up the development of the industry, CCSA TC610 has established an industry collaboration platform focusing on demand research, standard formulation, simulation platform construction, together with evaluation and certification programs promotion.

IETF has setup a working group for networking automation as ANIMA, working on the standardization focused investigations on intent driven networking as an enabler for service-based architecture.

ETSI ZSM has issued a number of autonomous network specifications, such as E2E network and business management architecture, closed-loop automation, service management, and sustainable development of a diverse set of services. It has also approved projects for intent interface, AI enabling under cross-layer and cross-domain architecture and digital twins, fulfilling the shortcomings of the traditional human centric network management along with the need to support software-defined, flexible and service based architectures.

ETSI ENI was established in February 2017, which is committed to defining the cognitive network management architecture, and adjusting the services provided according to user needs, environmental conditions and changes in business objectives through the use of AI technology and context aware policy. At present, the organization has carried out intent policy[29-31], knowledge graph[32-33] and other key enabling technologies related to intelligent networks. The Group's work encompasses Artificial Intelligence (AI) capabilities in a closed-loop control mechanism to improve performance of network orchestration and resource management functionalities. The intelligence is offered to the emerging technologies like network slicing, SDN and NFV for enhancement of their operational and maintenance lifecycles.

2.4 Gaps and Challenges facing highly autonomy

As more and more CSP making plans, setting goals and promoting large scale deployments of AN, different challenges keep surfacing per domain, per service and per CSP. In the 1-to-N process of AN realization and application innovation, the industry are currently facing some common challenges expanded in the following areas.

- Achieve minimal OAM under the continuous growth of network complexity and business diversity. It's impossible for traditional linear improvement of human OAM efficiency to cope with the exponential growth of complexity added up by expansion of mobile communication spectrum utilization, diversification of user scenarios and traffic, and multiple antennas in RAN domain. The complexity of these networks also increases due to the introduction of virtualization and containerization technologies to networks also brought great challenges.
- Lacking of common agreement on L4 profile and evolution path. Though the engaged parties of autonomous network deployment have reached on a very rough idea of L1-L5 evolution path and roadmap, there are still undeniable differences when it comes to the profile definition and measurement on L4 autonomy. Based

on the first technical maturity definition come up in 2020, L4 refers that all the execution, awareness, analysis and decision tasks are accomplished automatically by telecom system itself. And intent management tasks can be partly accomplished automatically by telecom system itself based on human defined intent translation policies. However, it can be interpreted by different parties in various ways and angles, especially when it comes to L4 standard for each network domain. For example, China Mobile targets a L3 autonomy of automatic implementation ruled by dynamically programmable policies, and L4 autonomy of automatic implementation ruled by AI assisted knowledge, featuring with continuous learning and rapid evolution. While Google and Vodafone define a path from L2 workflow-driven, to L3 event-driven, and to L4 machine-driven.

- Coordinated evolution vision of autonomous network towards 5G-A supporting. In 2021, the standardization of 5G-Advanced was officially launched in 3GPP. Expected to guide the network development after 2025, 5G-A is positioned as the core infrastructure of digital intelligent society. 3 major visions have been put forward as "excellent network", "intelligent and simple" and "low-carbon and high-efficiency"[34], which are supposed to comprehensively deepen and enable the transformation of digital intelligent society. The current industrial exploration on 5G-A innovation clearly put new and higher requirements for collaborative evolution of autonomous networks, since L4 autonomy by 2025 has been a common target for a great number of leading CSPs.

- Computing Force Network provides whole new business model and scenario. The essence of computing force network is that CSPs satisfy the business requirements and needs of cloudification of enterprises. Computing force, storage and network resources are allocated and orchestrated flexibly among cloud, edge and end as needed, which puts forward higher requirements for the automation and intelligent scheduling of operator network and cloud service. How could L4 autonomous network enable computing force network, the hottest high-value business scenario in the industry, has become a trending topic.
- Energy saving with network performance assurance. In 2030, DOU will reach 600G. The traffic of whole mobile network is expected to increase by 100 time, driving the number of base stations, frequency spectrum and channels to continue to overlap, resulting in a substantial increase in wireless network energy consumption. CSPs all over the world have put forward various energy-saving strategies, including VDF, Deutsche Telekom, China Mobile, China Unicom and China Telecom.

Being the necessary path to lead communication network to digital, intelligent and green development, L4 autonomous network are supposed to solve the above challenges in next 3 years.

3. Emerging technologies towards L4

To cope with above challenges, it is necessary to further integrate intelligence into network experience improvement and efficient operation and maintenance. Numerous emerging technologies has been hotly discussed and tested to do so. This chapter has selected MLOps, knowledge management and digital twin network, which tackles the challenges emerged during large-scale deployment and enables L4 and L5 AN in the following ways respectively.

- Model lifecycle management after large number intelligent network applications deployment. With large scale implementation of autonomous network, various intelligent applications are being deployed on various management layers and domains (e.g. network elements, network element management, domain management, service management, etc.) in a large-scale and distributed manner, which brings the problems of deployment and maintenance of AI/ML models used by those applications.
- Automated and intelligent management of AN knowledge. A large number of valuable expert experiences, rules and policies generated during the evolution from L2 to L3 are regarded as new-type "knowledge assets", the management, updating of which need to be more streamlined, automated and intelligent in L4 autonomous network.
- Decision intelligence towards L5 fully autonomy. Fully trusted and fully authorized decision intelligence is the premise for the network to be free of manual dependance and move towards L5 fully autonomy, which requires a highly efficient and low-cost testing ground and closed loop optimization mechanism provided only by the digital twin network.

3.1 AI/ML Operations (MLOps)

MLOps is a management practice for machine learning models, aiming to link up the development, deployment and operation of AI models, connect algorithms, business and operation teams, improve the efficiency of life cycle management of AI models, and promote their large-scale application. It can be seen as an extension of DevOps methodology to deliver machine learning models instead of software.

There has been open discussion on the importance of integrated ML life cycle management in industrial production in 2018. And it was then that foreign head enterprises starting to establish MLOps system. Since 2019, MLOps has been listed in Gartner's data science and machine learning technology maturity curve for two consecutive years, and is regarded as an important part of AI engineering. In 2021, XOps, including MLOps, was listed by Gartner as one of the top ten data and analysis technology trends in 2021. In 2022, AI engineering was listed as one of the key strategic technology trends by Gartner. IDC predicts that by 2024, 60% of Chinese enterprises will operate their ML workflow through MLOps. According to Cognlytica, the market size of MLOps will increase rapidly from 350 million USD in 2019 to 4 billion USD in 2025. It can be predicted that MLOps, as part of AI engineering, will be more widely applied and developed in the next two to five years.

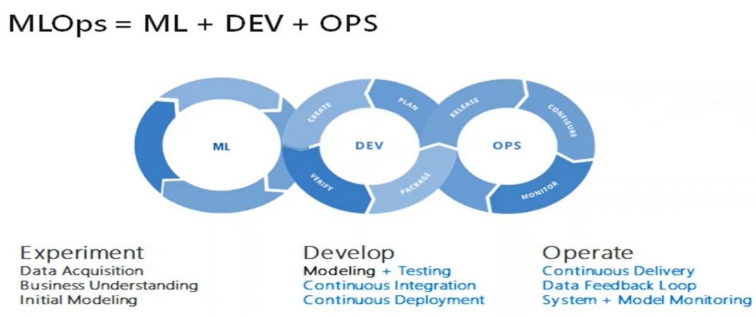

Figure 1 MLOps

MLOps includes 3 steps in general.

1) Project design, including requirements collection, scenario design, data usability check, etc.

2) Model development, including data engineering, model engineering, evaluation and verification, etc.

3) Model operation, including model deployment, CI/CD/CT workflow, monitoring and scheduling, etc.

3.1.1 Vision

With the large-scale implementation of intelligent network applications, the problems faced by its going further towards highly or fully autonomous level are shifting from model and algorithm-oriented problems to engineering problems. As an enabling technology of AI engineering, MLOps can effectively solve the practical problems of autonomous network application, including intelligent network management, intelligent operation, and intelligent network element management, improving research and development efficiency of intelligent network applications through systematic and automated lifecycle management of applications.

MLOps could be applied to enable the following autonomous network scenarios.

- Continuously monitor the effect of autonomous network applications, avoid the risk of model degradation caused by data drift, and support iterative training and continuous optimization of the model.
- Standardize the workflow of autonomous network application from model development, model delivery and model operation, and improve the quality and efficiency of model delivery.

The development of MLOps is directly driven by multiple challenges in life cycle of AI models from R&D to deployment.

- Automation: Everything that can be automated in the whole workflow should be automated, from the data access to the final deployment and launch.
- Continuous：MLOps adds the concept of continuous training on the basis of continuous integration, continuous deployment and continuous monitoring, that is, the model can be automatically trained and updated during online operation.
- Versioning: Version management is also one of the best practices of DevOps. In the field of MLOps, besides the pipeline codes, version management of data and models is emerging demands, which also poses new challenges to the underlying infrastructure.
- Experiment Tracking: experiment management can be taken as the enhancement

of commit message in version control. For code changes related to model construction, the corresponding data, code versions and corresponding model artifacts should be recorded and archived, which can be used as an important reference for subsequent model analysis and decision making.

- Testing: Machine learning system mainly involves three different pipelines, namely data pipeline, model pipeline and application pipeline (similar to the integration of model and application system). For these three, it is necessary to build corresponding data feature tests, model tests and application infra tests to ensure that the output of the whole system is consistent with the expected business objectives.
- Monitoring: In addition to traditional system monitoring, such as logs, system resources, etc., machine learning systems also need to monitor input and output data and model-related indicators to ensure the quality and efficiency of prediction, and automatically trigger coping mechanisms in case of anomaly, such as data or model degradation, model retraining and deployment, etc.
- Reproducibility: Different from the deterministic behavior of traditional software systems, there are many "random elements" in machine learning, which pose certain challenges to the troubleshooting of various problems, version rollback and the certainty of output effect.

3.1.2 Outlook

The application system of autonomous network is developed in a hierarchical manner. For the AI model management of network operation and maintenance level, CSPs tend to build their own AI platform, evolving from computing platform and application hosting platform to an MLOps platform integrating visual modeling, model management and model operation. A series of standards on "Intelligent Operation and Management of communication network " has been proposed in CCSA, in order to define a general AI engine for network operation and management level.

However, at the level of specific network domain and network element management, there is a lack of industry consensus on AI model management embedded in software and hardware products of vendors. At present, the relevant work only involves model training, and fails to cover the whole process of MLOps (the E2E automated scheduling, deployment, monitor, update, test and delivery of relevant data, models and applications).

As the objective of autonomous network is to replace the manual work of developing, installing, deploying, managing, optimizing and retiring network functions, it is certain to have a significant impact on the way that the LCM of network software works. Specifically, as AI/ML has proven to be an efficient tool to develop functionality for autonomous network, different options for training and inference of ML models will drive corresponding options for the LCM of software with AI/ML-based network elements

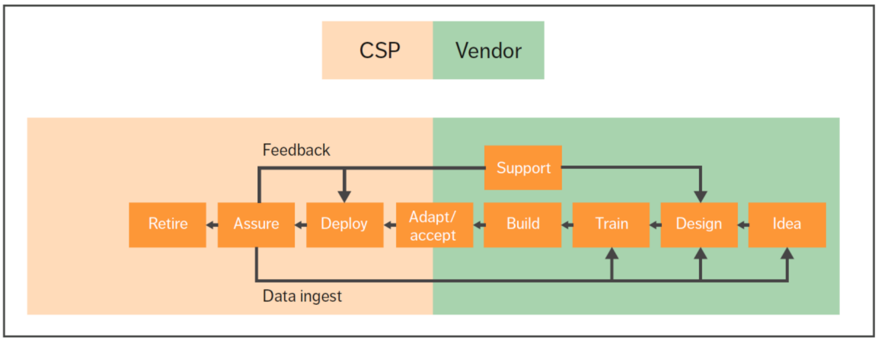

Figure 2 A high level view of LCM process

Figure 5 presents a process view of the LCM of RAN components, ranging from the initial idea for a RAN component to its eventual retirement. A RAN component is defined as either a pure software entity or a hardware/software (physical network function) entity.

An important aspect of the LCM is that it represents a structure of responsibility and ownership among vendors and CSPs. This structure is the baseline for the business model between vendors and CSPs, structuring exactly what is delivered by the vendor in the LCM process. The light orange and green background colors in Figure x highlight the responsibilities of the CSP and the vendor respectively. Software or software/hardware entities are delivered in the adapt/accept step together with support contracts and, in some cases, professional services for integration and deployment.

Using AI/ML models in the RAN automation solution requires the introduction of a model training step to the LCM process. There are four main alternatives for how to add model training to the LCM, each with implications on the responsibility split between the vendor and the CSP.

i. The first alternative is for the vendor to deliver a global model (that is, the same model for all CSPs) in the form of software entities in the adapt/accept step. A global model can, for some use cases, still allow for consideration of local context and can be very powerful in creating highly flexible automation functionality that can adapt to different deployments. In this case, all training is the responsibility of the vendor and occurs in the train step.

ii. The second alternative is for the vendor to deliver local models in the form of software entities tailored for different uses (CSP-specific or geo-specific, for example) in the adapt/accept step. Local training is the responsibility of the vendor and occurs in the train step. This full model training alternative requires access to local data, and it is important to be aware that the cost of maintaining different software versions could become substantial. As a result, this alternative is most appropriate for scenarios with centralized inference in a few places per CSP where there is only one or just a few ML models that do not require frequent retraining. In scenarios with distributed inference in thousands of places per CSP that require retraining every other week (for example), this model training would not be the best alternative.

iii. The third alternative is for the vendor to deliver a global model that can be retrained on additional data sets. In the adapt/accept step, the vendor delivers the model in the form of software entities together with information about how to retrain and evaluate it. The CSP is responsible for retraining the model to become a set of local models, which expands the adapt/accept step to include training. In these scenarios, it is unclear how much responsibility the vendor can take for in-field performance and support. Therefore, there will be practical challenges to use alternative iii in commercial deployment until responsibilities have been resolved.

iv. The fourth alternative is for the vendor to deliver a base-trained model in the form of software that is designed to be automatically retrained on local data after deployment. We refer to this as embedded training, and the training is transparent to the CSP. In this case, the training is the responsibility of the vendor and occurs both in the train step and autonomously in the deployed software. This is a path toward a fully autonomous system, while keeping the current business relation between vendor and CSP intact.

One of the key principles to be mentioned is, if the model training/retraining needs to use data from CSP network (as in most of the cases), some kind of data usage agreement between vendors and CSPs should be reached as a precondition. But also in most cases, there is a lack of access to data and the management of access to real data due to regulations regarding privacy.

For above four ML model training alternatives, alternative i and iv are already used in commercial network and generally accepted by the industry since they have clear responsibility boundary and are similar to traditional LCM of RAN applications And from model training data usage perspective, alternative i is offline training, data from CSP network should be provided to vendors' domain for model training; For alternative iv, it is online training, model training/retraining process is automatically executed in CSP network; In such case, data of CSP network for model training/retraining are not needed to transfer to vendor's domain.

Alternative ii and iii also have potential to be used in real deployment. For alternative ii, the responsibility boundary between vendor and CSP is clear. The only thing to be aware is that the efforts and cost to maintain different software versions, so the centralized ML inference scenario with ML models that not requiring frequent retraining should be suitable for alternative ii. For alternative iii, it gives CSP the responsibility and flexibility to train local modes, but in practice, it is not clear how much responsibility the vendor should take for field performance and support. Therefore, it should be further evaluated and observed in the near future if alternative iii can be adopted in commercial deployment.

3.1.3 Approach

Compared to LCM of traditional software used in network, AI/ML introduces the elements of model training, inference, a stronger need for data sharing and relevant data access technologies with privacy protection. It is proposed that

- Enhance standardization on LCM alternatives on key scenarios. The industry must adjust the LCM of AN functionalities to include AI/ ML-based technology to reach its potential as it evolves. It's important that the industry agree, as much as possible, on the LCM architecture and alternatives on vital scenarios, take the resulting LCM processes as the baseline for AI/ML-related standardization. CSPs and vendors should work closely in SDOs to avoid extra cost caused by unclear responsibility segmentation and industry fragmentation.

- Promoting standard interfaces and algorithms creation in open source communities. Open source has moved from creating technology that can be used to build networks according to standards to create ecosystems around default interfaces. In this case, it could be used to speed up maturity of AI/ML technology, platforms and tools, provoke innovation for different inference and training techniques, and realize reference of standards interfaces.
- Jointly work on research and adoption of Privacy-Preserving Computing. Privacy-preserving computing technologies, such as federated learning, have developed rapidly in recent years. There are still difficulties existing in its large-scale application, including performance bottlenecks, security certification, poor standardization of data quality and fairness problem. It is necessary for CSPs and vendors to cooperate on consolidating the theoretical research, carrying out engineering exploration and solving data island problem, constantly improving the technology maturity and industrialization ability in practice, enhancing the value of data collaboration, and promoting the implementation of LCM alternatives.

3.2 Knowledge Management

Knowledge is the concept, notion or skill acquired through study, practice or exploration. [35] In autonomous network, it can be expert experiences, policies, rules, etc. Knowledge management is the life cycle management of knowledge, including knowledge construction, knowledge processing, knowledge sharing, knowledge application, knowledge updating, etc.

3.2.1 Vision

Autonomous network knowledge is applied to the whole life cycle management of autonomous network (planning, deployment, maintenance, optimization and operation) to enable system automation and intelligence. It can be embodied in pre-configured static rules, programmable dynamic policies, etc. which effectively drive the close loops for automation.

For highly autonomous network, where the dynamically changing network/system has brought complex and diverse knowledge management requirements. The automated and intelligent management and operation of knowledge has become a new target.

Advanced knowledge management system (as separated from simple execution engines for rules/policies) is supposed to improve the quality of generated knowledge and promote efficient lifecycle operation of knowledge, so as to achieve timely update and accurate application of knowledge. At the same time, through cross-domain knowledge creation, integration and transmission, knowledge sharing and collaboration at different layers and levels can be realized. Redundant processes can be further streamlined, automated management processes fulfilled, operation efficiency improved and management costs reduced, especially in the situation where manpower can't meet the huge demand of knowledge management. Knowledge management cooperates with autonomous network evolving, which is applicable to the phased evolution path of autonomous network, featuring informatization, automation and intelligence.

3.2.2 Outlook

It has become a new demand for knowledge management in the digital era that how to efficiently match the constantly changing knowledge resources to the dynamic knowledge demand, so that knowledge can be self-iterated and upgraded, and generate value in practice. In TM Forum and ETSI, preliminary research has been conducted. It mainly focuses on the requirements, challenges, reference points of architecture and definitions of knowledge management and applications:

- The sub-domain of Knowledge Management in eTOM and TAM standard projects of TM Forum has firstly combed knowledge management from the perspective of business and application. The ODA project team released IG1130F in June, 2019 to study the requirements of knowledge management applications, and proposed modifying suggestions for knowledge management applications in TAM. The AIOps project team released IG1190E in May, 2020, and studied the challenges and opportunities faced by knowledge management practice after the introduction of AI, as well as the new process principles, etc. The AN project team released IG1251 in July, 2021, studying the autonomous network reference architecture, in which a cross-domain module naming “knowledge and intelligence” and a single-domain module “domain intelligence” were defined, providing knowledge management and empowerment functions.
- In September, 2019, GS ENI 005 released by ETSI gave the definition of knowledge management function block of ENI system, and introduced the driving force, functions, operation process, etc. of knowledge management module. The draft GR ENI 015 studies the knowledge management of intent policy, and adopts the way of knowledge graph to manage intention policy. GR ENI 031, kicked off in March 2022, plans to study construction and application of fault maintenance network knowledge graphs.

The automatic and intelligent evolution of knowledge management introduces the following new challenges and opportunities:

The tooling for data acquisition needs improvements. Data is the starting point of knowledge construction. In practical applications, the diversity of data sources results in inconsistent data standards and poor data quality, resulting in multi-source data ambiguity, high noise, and unclear relationship between data. From the perspective of source form, knowledge is contained in structured (e.g. alarms, indicators, etc.), semi-structured (e.g. configuration, log, standardized product documentation), unstructured (e.g. practice manual, failure case, experience sharing, packet capture data on production network for an alarm failure diagnosis, etc.) data, even in the minds of experts. Correspondingly, we need to match the tools to obtain these data and provide "standard and understandable" specifications to allow interoperability. At present, both the tools for non-institutionalized data acquisition and industry standards need improvements.

The difficulty of knowledge extraction is increasing. With the application of AI technologies such as biometrics and object recognition, databases including fingerprint database, face database and image database are established. Knowledge extraction for unstructured data will no longer be limited to text, which poses higher challenges to unstructured data extraction.

The generally-applicable evaluation index is missing. Every single link of knowledge management, including knowledge extraction, modeling, fusion, inference, etc., involves different algorithms. However, the low generalization ability, low robustness and lack of unified evaluation index of algorithms in this field also add up certain obstacles.

Deployment method for knowledge collaborative management is needed. Autonomous network involves multiple business fields and heterogeneous networks. It is necessary to manage knowledge at different layers through cross-field knowledge creation, integration, sharing and collaboration. Considering the hierarchical characteristics of the network and management system, and the differences in IT resource requirements at each stage of AI-driven knowledge generation (knowledge acquisition, training, reasoning, etc.), it is necessary to build distributed AI capabilities in the network to support automatic closed-loop at each layer, as well as knowledge sharing and collaboration between different layers.

Knowledge management practices in autonomous network are still in an early stage with standardization going up front. But Knowledge graph, as one of the most effective and ubiquitous tools of knowledge management in the digital age, has already been widely applied in various domains. Knowledge is considered to constitute the foundation of most modern data and analysis capabilities. It is a semantic knowledge base with a directed graph structure, which is used to describe concepts and their relationships of physical world in symbolic form, transform internet information into a form closer to human cognition, and provide a better ability to organize, manage and understand massive information on the Internet. It’s the basis of knowledge-based intelligent services on the Internet, an important path to promote AI from perceptual intelligence to cognitive intelligence. In recent years, it has been widely used in intelligent search, intelligent question and answer, personalized recommendation and human-machine interactive dialogue, etc. The application of knowledge graph in telecoms gives full play to its advantages in cognitive understanding and intelligent analysis, formulating an intelligent brain for business and operation, and providing a data analysis engine for dynamic decision-making by using the cognitive relationship network of knowledge graph, thus supporting relevant operation scenarios, promoting the implementation of network operation scheme, and improving the service quality and operation efficiency. It's applied in CSP's network in the following scenarios:

- Knowledge graph of work sheet analysis: It is mainly used in the intelligent analysis scenario of complaint work sheets. By the entity extraction and semantic understanding of the work sheet content, the knowledge graph is formed, and the solutions are merged to form processing suggestions, which are automatically pushed to the handler for decision-making assistance, realizing the analysis and processing in a "man-machine" collaborative way.
- Question-answer retrieval knowledge graph: It is mainly used in intelligent response scenarios. By classifying and identifying user's questions, word segmentation and semantic analysis, key semantic is extracted to match the template knowledge graph. The entity or attribute content of the graph is obtained. The answers are queried in the graph database, and responses are formed.
- Knowledge graph of Alarm correlation: It is mainly used for alarm root cause identification. By mining all kinds of data (historical alarm, topology, application call relationship, configuration, etc.), the alarm correlation map is dynamically constructed, the recommendation of fault root cause correlation analysis results is realized.
- Knowledge graph of fault diagnosis: It is applied to network fault treatment scenarios. Firstly, the network fault knowledge graph is built on existing expert experience, fault treatment cases and logs, etc., through knowledge extraction, knowledge fusion and other steps. Combined with knowledge inference and fault diagnosis algorithms, intelligent network troubleshooting is realized.

3.2.3 Approach

The standards research of autonomous network knowledge management is still in a very initial stage. Leading the formulation of knowledge management standards in the context of autonomous networks, helping to guide the industry to reach a consensus on knowledge management, cross-domain knowledge sharing and network intelligent application, enabling knowledge management facilitates the hierarchical evolution of network autonomy development, and further promotes the standardization of existing network operation and maintenance processes.

In particular, three parties would be playing essential roles in the way forward:

- **Standards** for data collection, ontology models, knowledge representation as well as general functional architecture to establish a unified mindset around collaborative endeavors of integrating knowledge management into autonomous networks.
- **Practice** in production from service providers by introducing into the existing network operation and maintenance process to guide the application of and feedback into further research and development of systems or tools for knowledge introduction, life cycle management, and cross-domain knowledge sharing.

- **Implementation** from network element/network management manufacturers to decouple domain knowledge management and business logic implementation, hence providing flexibility in reference implementation for centralized and cross-vendor knowledge sharing and applications.

3.3 Digital Twin Network

The concept of digital twin has been in a constant evolution since it was first put forward in 2003. According to digital twin consortium, digital twin is a virtual representation of entities and processes in real world, synchronized with a certain frequency and fidelity. Digital twin was first used to model expensive assets or complex processes. By constructing digital twin, combining both virtual and real interactive feedback, data analysis, iterative decision optimization and other means, new capabilities can be expanded for physical entities. With the growing of IoT, artificial intelligence and other technologies, the research of digital twins has gradually extended to a broader field. Digital twin network came as a result of applying digital twin in communication industry.

The digital twin network is a virtual twin of the physical network entities created in a digital way, and can be mapped and interact in real time with the physical network entities. On the one hand, towards L4/L5 highly autonomous network, all the tasks of execution, perception, analysis and decision-making are supposed to be undertaken by the system, which poses higher requirements on automotive and intelligent level. On the other hand, the introduction of 5G and new services adds heavily on network scale and traffic load. In addition, the complex and diverse application scenarios have also brought great challenges to the mathematical modeling of the network environment and diversified target solving. Digital twin provides a good "testing ground" for autonomous network. It has the potential to change the established process of "planning, deployment, maintenance, optimization and operation" of network, and lay a solid foundation for the future L5 fully autonomous network. Through the real-time interaction between the physical network and the twin network, various states of network can be clearly viewed or played back on the digital twin network platform. Feedbacks after virtual deduction by network simulating and forecasting are produced. And parallel verification for multiple digital twin examples with different optimization objectives or innovative applications can be done without possibly affecting the existing network, helping to achieve low-cost trial, intelligent decision-making and high-efficiency innovation, and speeding up product R&D for network.

3.3.1 Vision

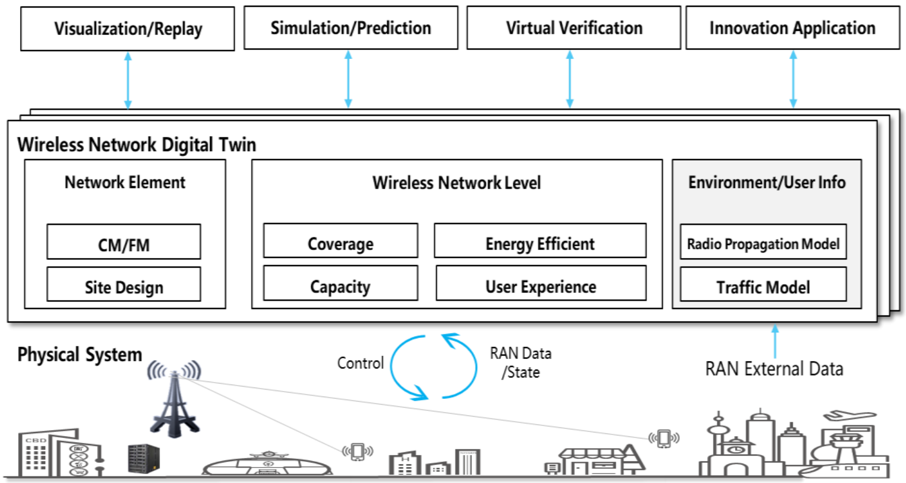

Figure 3 vision of digital twin network

Categorized by the modeling objects, there are 3 major directions regarding digital twin network applications.

- Digital modeling for physical network elements: the digitization of physical objects is relatively slow-changing, including the site design and configuration/fault status of physical network elements. The former is the description of hardware composition, topological connection and soft capability of physical sites, while the latter is the digital expression of network element status.
- External environment/user-oriented digital modeling: this is the digital depiction of the relatively dynamic user traffic and its communication environment. User experience of mobile communication is closely related to the trajectory of human activities. For example, the traffic model in the same area at different times is dynamic. And user's environment is also constantly changing with the flow of users. The application effect of digital twin depends directly on if we can accurately describe the influencing factors of these dynamic changes. The existing RAN interface can only obtain very limited information about the propagation environment and service model, so it is necessary to introduce external RAN data to describe it together.
- Digital modeling for RAN: An excellent radio access network achieves the best balance of 3C1Q (coverage, capacity, cost and experience). With the increase of wireless stations and carriers, energy consumption has also become an important evaluation index. Based on the information of network elements and external environment/users, considering a number of local information, and finally integrating into a complete RAN model that covers multiple dimensions of capacity, coverage, experience and energy consumption, etc.

Following network lifecycle management, the target application scenarios of digital twin network mainly focus on the following aspects.

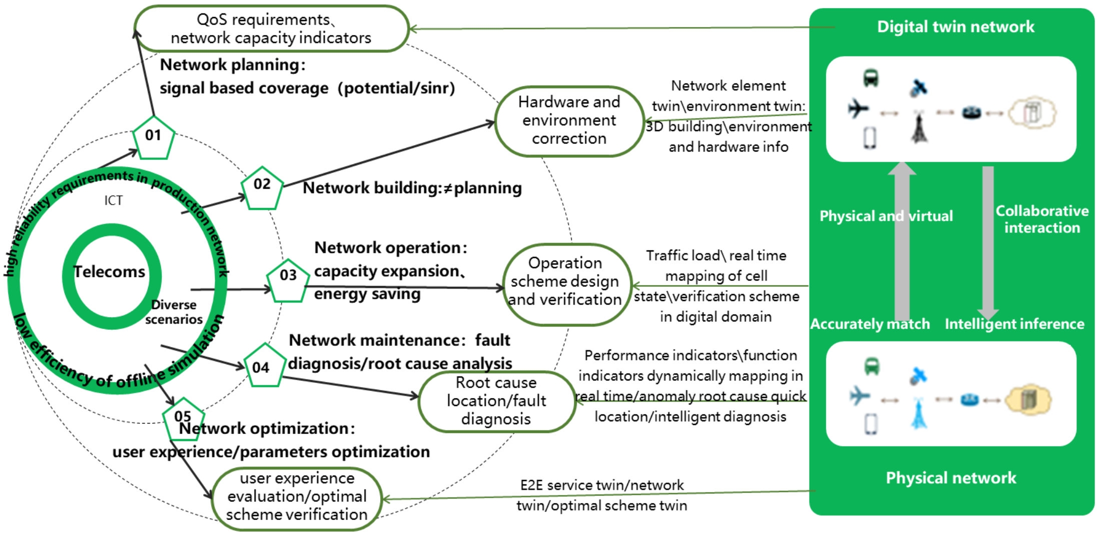

Figure 4 Digital twin in AN lifecycle management

- Network planning: enhanced user business simulation, enhanced environment 3D modeling.
- Network development: E2E project improvement tracking, connecting the network lifecycle to ensure the accuracy of data.
- Network maintenance: fault recurrence and scenario backtracking, help to locate and solve problems.
- Network optimization: support reinforcement learning to realize parameter optimization.
- Network operation: Evaluation of network capacity, performance and energy consumption, supporting the design and verification of operation scheme.

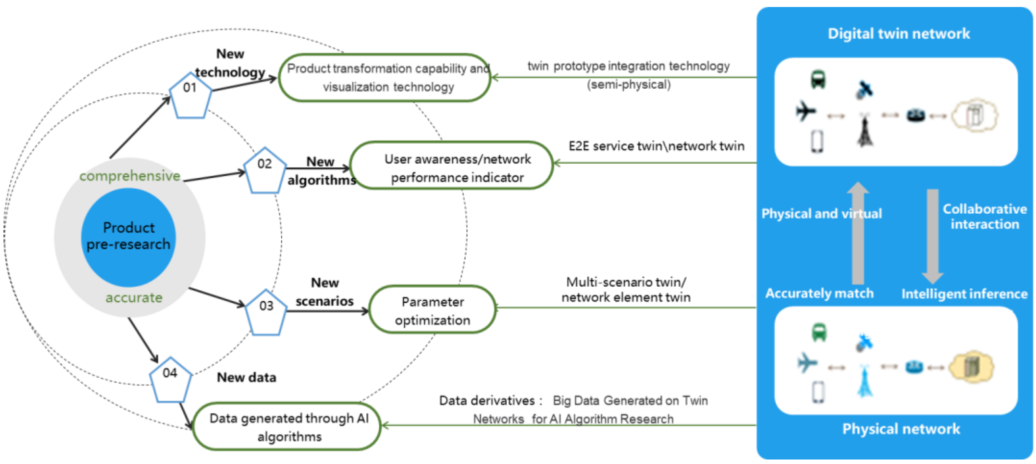

Figure 5 pre-research acceleration

Enabling ICT technology innovation, digital twin network can accelerate product pre-research in the following aspects:

- New technology: First of all, new hardware and software technologies can be realized efficiently and at a lower cost on the twin network to support the initial feasibility and value demonstration. Then, it can support the demonstration of new technology value combined with visualization technology. Based on the twin network, the prototype development can be completed efficiently, and the conversion of new technology products can be supported.

- New algorithm: accumulate outfield model, assist algorithm verification and accelerate version iteration.
- New scenario: To control the cost of outfield operation and maintenance, a one-size-fits-all default scheme is applied. Parameter setting needs to consider richer scenario models, and new outfield models could be generated based on the combination of existing outfield models to improve verification.
- New data: Data generated by AI algorithm. User scenario is customized according to requirements. Data is collected, cleaned and labelled automatically.

3.3.2 Outlook

Digital twin network has been a hot focus in the last 2 years, but the whole industry, on general, is still in a quiet primary stage of target applying scenario exploration and PoC. There are no mature products so far in the industry. The trade-off between cost and value has been the major bottleneck restricting its monetization. Digital twin is a methodology. Whether it being used or not depends on the true value that it can bring or can be expected to bring. There is no upper limit to depict RAN, since it’s ubiquitous anytime and anywhere. There is a trade-off between the cost being paid and the value brought by it, including the following aspects:

- Challenge of data: The existing data is collected at a pretty rough granularity and mainly used for network operation and maintenance. To build digital twins for network self-maintenance and self-optimization, it is necessary to depict the network based on massive data. This also puts much higher requirements on data diversity, accuracy, data distribution and sample labels. During massive RAN data collection, there are more problems such as resource consumption of embedded system, accuracy assurance of collected data, bandwidth of transmission channel, delay, data processing, storage cost and lifecycle management of data, which still need to be solved by technical means. Moreover, the acquisition of external RAN data is still facing the problem of lacking standardization, which can be explored in combination with specific cases.
- Challenge of modeling: How to model in layers and domains, and organically combine them to describe the whole complex RAN system pose a grand problem. Data is always limited so that migration and generalization need to be considered, that is, how to ensure the migration effect of the model based on limited samples. In addition, how to evaluate and ensure the accuracy of the model to meet the requirements of network self-maintenance and self-optimization requires a balance between efficiency and cost.
- Challenge of computing power: If the physical RAN is directly mirrored in equal ratio without any compression, the overall computing power demand will be unacceptable. And it will increase proportionally if real-time requirement being considered. It is necessary to use proper modeling and computing force contraction

techniques, and strike a balance between force demand conduction and modeling accuracy.

3.3.3 Approach

The concept of digital twin is constantly evolving with the related research and application going deeper and broader. Compared with previous exploration of GE, IBM and Microsoft in digital twin architecture and modeling language, digital twin network can only be regarded as just a new born.

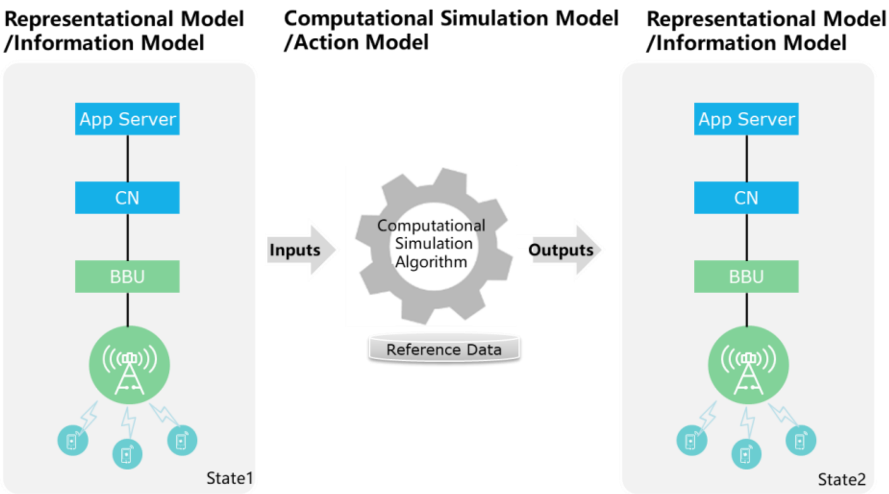

Figure 6 digital twin models

Referring to Digital Twin Consortium, digital twin models can be divided into two categories: representation model and simulation model. A representation model is a description of an entity or a process state, which can be understood as an information model while simulation model is an executable model oriented for process, which consists of input and output data and algorithms, and can be understood as behavior model. In this way, RAN digital twin models can also be divided into these two categories: one is the representation model for digital expression of RAN, and the other is the simulation model for simulating network changes after internal or external environment changes. Simulation abilities to solve single domain problems have been accumulated during the development of applications in network operation lifecycle in varying degrees. However, for the representation model, the existing data are mainly used for network operation and maintenance. The comprehensive cognition of wireless network depends on the correlation analysis and judgment by engineers. Considering the cost and value, it is suggested to start from specific applications, analyze the required internal and external data of RAN, and build the representation model and simulation model of key influencing factors. With the increase of such pilot applications, the digital twin network technology will grow fast and robust.

4. **Conclusion and suggestion**

Looking forward, the development of Autonomous Network is to be combined with the computing force network strategy and roadmap of 5G-Advanced, giving full play to the industry influence, focusing on the goal of innovative open platform formulation and industrial collaboration promotion. Standardization, open source community and industry promotion needs further enhancement on top-design, service design and AI technology application. Following suggestions are proposed:

In short-term, focus on the key autonomous network capabilities. The industry needs to reach further agreement on technical maturity measurement of L1-L5 autonomous network, and evolution path. The AI platforms of CSP could be better supported by deeper exploration on knowledge management and MLOps, etc,. to promote the internal sharing of computing force, algorithms and datasets, and therefore, monetization by scale effect.

In mid-term, facing the collaborative evolution with 5.5G and 6G network technologies, an open cooperation platform is built to promote the close cooperation among business, academy and research Institute. The platform opens up service scenarios, platform capabilities, data, knowledge, simulation platforms, as well as evaluation and certification services, etc., and guarantee for the training and inference of network intelligent applications, the exploration of 6G native intelligence networks and intelligent computing force networks.

In long-term, with the vision of empowering thousands of vertical industries, Holistic AI will be practiced continuously on technology, products, deployment and eco-system building. China Mobile plans to take the lead in setting up industrial chain innovation carriers such as intelligent network joint R&D laboratory, network intelligent collaborative innovation base, etc., through demand traction, R&D cooperation, investment enhancement, etc., providing support and assurance for our partners to jointly define L4 target portrait and realize the zero-X and self-X vision.

5. Applications and Solutions

5.1 Use Case Analysis in Workflow Domain

5.1.1 Network Planning

RAN construction and expansion cannot be achieved without network planning. Traditional network planning, especially expansion planning, requires on-site drive test and survey, and analysis together with network management statistic data. The whole process is manual operation with high cost and long planning period, and it is difficult to achieve network-wide precise planning. As a result, the network often needs adjustment for many times or the budgets in some areas are insufficient, affecting the overall network quality and user experience.

With the introduction of big data and AI technology, accurate network-wide automation planning has become a reality. The precise planning of 4G is based on big data, automation, and AI technologies to collect MR and network performance data of users. Combined with field engineering parameter and map information, the area where is considered highly valuable but with insufficient user experience can be identified from multiple dimensions including network coverage, network capacity and traffic, etc. Sites are automatically planned for construction, matching appropriate site types. Then, models are built based on AI algorithms to predict the effect after base stations activation. Finally, the priority of site construction is sorted out and the planning report is automatically generated.

Based on “big data + automation +AI”, the precise network planning can fully consider all mattering factors from the network vision and generate accurate predictions, enabling the network planning more in line with the actual service development requirements. In this way, the coordinated planning of 4G/5G coverage and capacity can be realized to maximize the benefit of network construction investment.

5.1.2 Network Deployment

With 5G technology and standards going stable and CSPs’ large-scale 5G network deployment, mobile communication service has gone into a whole new era. In the early stage of 5G deployment, CSPs mostly focus on the deployment goal of completing the full coverage as soon as possible, to win advantages in business competition.

Basic parameter configuration is critical to site commissioning, network access and KPI acceptance criteria. As for new 5G sites, neighbor cell configuration, ENDC X2 and Xn configuration, as well as subsequent neighboring cell optimization and PCI optimization, are the first step of network commissioning and initial optimization. Inefficient manual operation has posed grand difficulties in large-scale network commissioning. For which 5G SON solution has been found efficient in locating the anomaly, improving maintenance efficiency, and enriching maintenance methods. It realizes intelligent identification, self-organization, orchestration and error correction of neighbor cells and links, helping the rapid construction of 5G network.

5.1.3 Network Maintenance

AI brings the most important solutions to deal with new challenges brought by the increasingly large and complex multi-layer and multi-standard wireless communication network.

For increasingly dense multi-layer networks, the traditional way is to check a single KPI each time through a variety of network optimization tools by an experienced engineer, then to combine one or more KPIs with the expert experiences for analysis. Together with pre-defined rules or threshold values, the degraded cell are firstly located, then analyzed with more relevant KPIs and MR information to find out the root cause, then generate the optimization solutions for deployment and verification. This traditional way of dealing with the problem requires even more engineers in the case of massive communities in large modern cities.

Using state-of-the-art machine learning algorithms to autonomously discover problems in massive data and quickly identify and classify them can significantly improve the network optimization efficiency. The transformation and skills upgrading of network optimization personnel and the application of AI modules make it possible to handle increasingly complex network problems.

In addition to the machine learning algorithm for network problem identification, classification and root cause analysis, it is further combined with the expert experiences in network optimization to carry out intelligent maintenance and optimization, which can further provide the root cause and solution of the degraded cells. For specific scenarios, such as high-loading scenarios, optimization solutions for specific parameters can be provided.

Another problem faced in reality is the trust issue and interpretability of AI. The complexity and sophistication of AI systems has skyrocketed to a point where humans do not fully understand the complex mechanisms by which AI systems work or how they make certain decisions, such as those using Deep Neural Networks systems, which are considered complex black-box models.

The inability for humans to see inside the black boxes may hinder the adoption of AI (and even its further development), which is why the urgent need for interpretability, transparency and understandability. Explainable Artificial Intelligence refers to methods and techniques for generating accurate, explainable models of why and how AI algorithms arrive at specific decisions so that humans can understand the results of AI solutions. XAI enhances trust in AI-based solutions by providing the needed understandability and transparency, elements that are critical to ensure that AI-based systems can be understood and trusted for long-term stable applications. How to better balance the performance and interpretability of AI models is also an area that the

industry needs to think about and explore.
Research on aspects such as trustworthy AI and XAI will require the collaboration of people from multiple disciplines to jointly advance the development of this field. It is envisioned that cognitive networks will be a central feature from 5G to 6G, which includes AI models being deployed at scale to raise the network intelligence. Due to the expected scale, complexity, and criticality of future networks, understanding and trusting these AI models and their behavior through seamlessly integrated XAI technologies becomes even more crucial.

Alarm handling is another important task of network O&M. There are many types of equipment in mobile network such as base station, transmission, power supply, etc. In addition, base station also has BBU and RRU, which are closely related to each other. A fault often generates a lot of non-root cause alarms. Quickly finding out the root alarm and fast troubleshooting are very helpful to build a high-quality network.

The intelligent alarm handling solution uses AI through big data mining and analysis to achieve automated correlation of alarms and to reduce their number. It can rapidly locate fault causes and present them through graphic UI. It can help operator greatly shorten time for alarm location and help network O&M reduce costs.

In the future, more alarm correlation rules will be dig out from big data of network to achieve a higher alarms reduction rate. Automatic diagnosis and self-healing capability will be achieved in case of faults, requiring no manual intervention.

5.1.4 Network Optimization

Massive MIMO, one of the key technologies of 5G, uses large-scale array antennas and 3D beamforming to effectively improve the coverage and system capacity in complex scenarios. Compared with traditional antennas, Massive MIMO with large-scale array antennas have more dimensions for parameter adjustment, including horizontal beam width, vertical beam width, azimuth, down tilt, and number of beams. Tens of thousands of possible antenna parameters could exist in one cell theoretically, which make it almost impossible to manually complete multi-cell collaborative optimization and adjustment for different scenarios or services.

Intelligent coverage optimization solution based on MR data and search algorithms achieves automatic data collection, optimization analysis, automatic parameters activation and verification, which greatly saves manpower and time, improves network optimization efficiency greatly. The solution can also self-adapt to identify and optimize scenarios with tidal effect or with sudden traffic changes.

5.2 Use case : Predictive Fault Management

5.2.1 Scenarios and Challenges

There are three levels of Intelligent wireless network fault management. The first level is automatic ticket processing, which significantly improves maintenance efficiency. The second level is proactive maintenance based on prediction and prevention technologies, which transforms the O&M mode from responsive to predictive. The third level involves providing a highly reliable self-sensing and self-healing mobile network that features intelligent O&M based on advanced perception capabilities and various technologies, including intent-driven and multi-layer collaboration.
Most operators are at the first level. Pioneering operators are developing automatic troubleshooting capabilities and shifting their focus to intelligent troubleshooting. Proactive maintenance has become the industry consensus. In recent years, leading operators have posed requirements on the scope of fault prediction and the depth of the fault prediction capability in terms of many aspects, ranging from strategic indicators to platform capability development. Wireless networks have been facing a range of issues, including many potential fault points, various fault symptoms, and complex fault causes (hardware, software, external system, environment, or human factors), fault pileup on active and passive devices (such as passive fronthaul, power and environment, and device faults), uneven fault detection methods or granularities, and outdated and complex onsite troubleshooting or inspection operations. These issues hinder the development of automatic and intelligent capabilities, including precisely identifying, demarcating and locating, troubleshooting and inspecting, and predicting wireless network faults.

5.2.2 Solution

- **Predictive and proactive O&M**

To cope with the preceding challenges and move towards autonomous networks, operators and manufacturers have carried out many innovative practices based on intelligent architecture. The fault detection capabilities have been enhanced, achieving comprehensive management. Multi-layer collaboration, cross-site collaboration, and device-network collaboration have been implemented to obtain optimal automation results.
In the high device temperature prediction solution, for example, the intelligent RAN architecture is used by NEs to detect the temperature data of each device in real time, predict the short-term temperature trend at the NE level, and report the trend to the RAN OSS.
Based on long-term data of multiple sites, the RAN OSS uses the neural network deep learning algorithm for modeling. After the baseline AI model is generated by training, it is generalized into site- and region-level models, which are then used for single-site inference. In this way, the optimal prediction performance is achieved and potential high temperature risks are identified in advance.

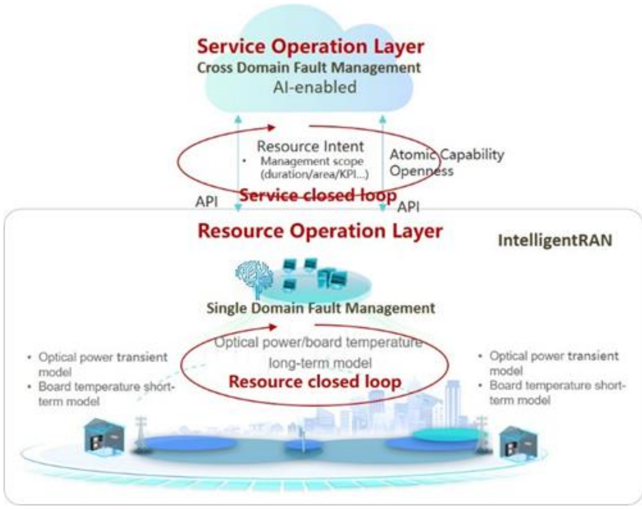

Figure 7 High device temperature prediction solution

The prediction capability can also be used for other scenarios, including identifying risks of optical modules and paths. Based on the long- and short-term perception data of the RAN OSS and NEs, the subhealth status of optical components and paths can be predicted.

- **Intelligent AR inspection**

In the "last mile" practice based on device-network collaboration, AR technology is applied to site devices. The image recognition technology and device-network collaboration enable automatic AR inspection, dumb resource modeling, and real-time fault point identification. A wizard-based intelligent troubleshooting engine is developed on the RAN OSS to communicate with site devices in real time and help site personnel perform efficient and accurate operations, greatly improving onsite troubleshooting efficiency.

Automatic equipment room inspection based on image recognition: The inspection scope covers 11 inspection categories and 25 sub-scenarios (such as equipment room door locks, batteries, and grounding surge protection), and 9 types of problems (such as battery rust and warranty period). Fast picture collection and equipment room inspection help engineers improve single-site inspection efficiency and ensure that all inspection items are checked.

Proactive risk rectification assisted by knowledge graphs: Inspection personnel check alarms and performance of devices based on inspection rules. Anomalies can be handled by referring to the online knowledge graph of the RAN OSS. Alternatively, inspection personnel can request expert support in one-click mode, greatly improving the efficiency of equipment room inspection and troubleshooting.

5.2.3 Benefits

The high temperature prediction solution has yielded satisfactory results in China (Anhui and Yunnan provinces) and Thailand. High board temperature problems can be predicted five days in advance with an accuracy of more than 90%. Possible causes and troubleshooting guidance are provided. Early troubleshooting significantly lowers the risk of base stations going out of service due to high temperatures.

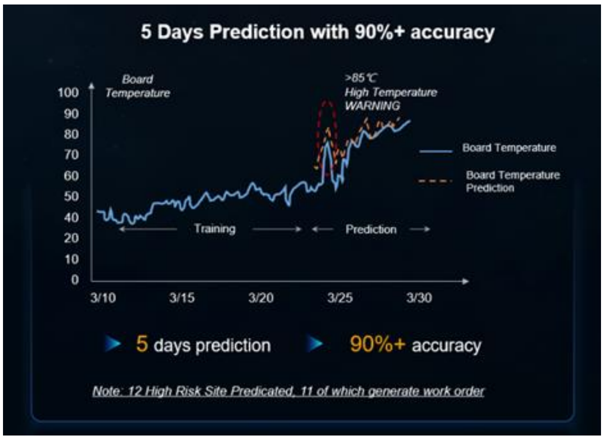

Figure 8　Test result of the high temperature prediction solution in China

The AR site inspection solution has been verified in many provinces and cities in China. The solution can be used to read almost all dashboards and items that need to be manually identified in wireless equipment rooms, shortening the inspection duration from 70 minutes to 30 minutes when compared with traditional methods. The AR-assisted troubleshooting prototype solution can reduce the mean time to repair (MTTR) by more than 50% in typical scenarios after onsite troubleshooting verification of fronthaul faults.

5.3 Use Case: Dynamic Quasi-Real-Time Optimization

5.3.1 Scenarios

On mobile networks, many factors, including the radio environment, traffic volume, network load, user locations, and service experience, vary over time. The frequencies of these changes are different. Examples are as follows:

- Monthly or weekly user growth
- Daily or hourly traffic changes
- Minutes-level emergencies
- Seconds- or milliseconds-level service experience deterioration

Figure 9 Dynamic changes of wireless network traffic over time

Dynamic network changes, especially those that happen within a few minutes or seconds, pose tremendous challenges to network O&M. When a network emergency occurs, traffic converges quickly, and there is a rapid increase in the usage of air interface resources, resulting in complaints from cell edge users due poor user experience. If such an event occurs, the experience of specific users is reduced to a level below the SLA, affecting the experience of the entire service. It is difficult for O&M personnel to quickly monitor and respond to network or experience changes that occur within a few minutes or seconds. These network changes are averaged in periodic and regional monitoring results and it is difficult to quickly perceive their impact on user experience. The entire process of data collection, root cause analysis, and closed-loop decision making may be delayed for more than one or two hours after a complaint is received.

To address these dynamic network changes, intelligent capabilities are introduced into mobile networks. Based on the intelligent RAN architecture, base stations interwork with the RAN OSS to quickly detect changes and adaptively adjust and optimize networks.

5.3.2 Solution

To adapt to dynamic network changes, an intelligent network needs to provide quasi-real-time optimization capabilities, including:

- Network-wide quasi-real-time perception and prediction of network or experience changes

- Quasi-real-time optimization and closed loop for traffic bursts or experience deterioration

The entire solution consists of three parts: time series modeling, network-wide quasi-real-time perception and prediction, and quasi-real-time optimization and closed loop.

KPI change feature extraction and time series modeling: Network KPIs, such as traffic, resource usage, and throughput of each cell change over time. Network data within a certain period can be collected to extract the features of long-term and short-term changes in order to build a cell-specific time series model that can accurately predict KPI changes.

Network-wide quasi-real-time perception and prediction of network or experience changes: The time series model for network KPI changes and computing power can be leveraged to automatically identify traffic bursts or experience deterioration on the entire network. The identification result can be immediately reported to the upper-layer management and control node. Compared with traditional network monitoring, intelligent collaboration at the network and NE levels in the intelligent architecture achieves minutes-level and seconds-level quasi-real-time identification. Based on the short-term trend, KPI trends can be predicted to identify network or experience risks.

Quasi-real-time optimization and closed loop for traffic bursts or experience deterioration: Problematic cells and correlated cells are quickly identified to optimize parameters (including MLB parameters, multi-frequency camping and mobility parameters, RF beam parameters, and power control scheduling parameters) based on the topology, scenario analysis, and model technologies.

5.3.3 Benefits

In Tianjin, shopping malls and pedestrian streets are located in business districts, and residential areas are located in adjacent districts. Traffic is concentrated in different areas in different time segments (shopping malls and pedestrian streets during the day, and residential areas at night). Based on the traffic changes predicted by the time series model, SSB beams of 5G cells were dynamically adjusted to ensure stable KPIs even with a 5 to 10% increase in the traffic and number of users.

In the university and industrial areas in Shandong province, traffic changes dynamically in industrial parks, urban villages, hospitals, residential areas, and colleges in different periods. Dynamic optimization of SSB beams in 5G cells ensures stable KPIs even with a 5 to 10%+ increase in the traffic and number of users.

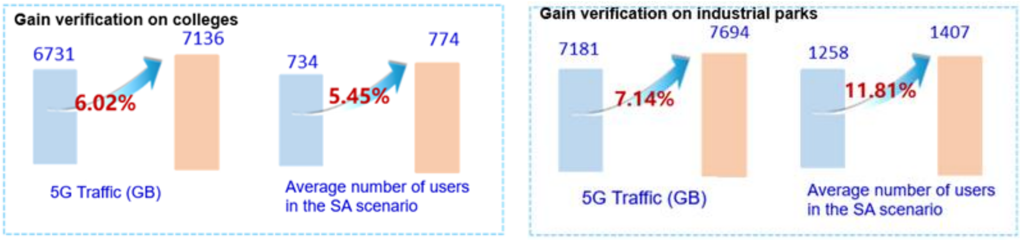

Figure: Verification of dynamic quasi-real-time optimization in different scenarios and periods

5.4 User Case: SLA-oriented Precise Network Planning and Assurance

5.4.1 Scenarios

5G networks are responsible for enabling digital transformation of industries. In recent years, 5G applications have developed rapidly in various industries, such as coal mines, ports, steel, manufacturing, and power grids. These industries are sensitive to costs and require SLA assurance (including network rate, latency, and packet loss rate). 5GtoB network O&M faces the following challenges:

- Intelligent network planning is required to provide network coverage and resource assurance based on SLA requirements.
- Precise service provisioning is required to reduce site survey costs and speed up service rollouts.
- User-level real-time monitoring and fault demarcation and locating must be performed to improve O&M efficiency and ensure SLA assurance.

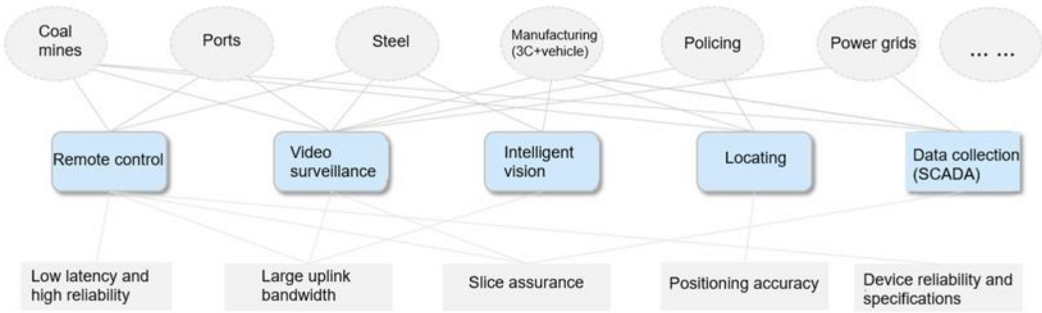

Figure: Multiple factors affecting 5GtoB SLA

5.4.2 Technology Trend

Due to the variety of 5GtoB services, high SLA requirements, and complex application environments, the traditional network planning mode based on expert experience cannot meet the deterministic network planning requirements of industries. The AI-based intelligent network planning platform is important for the commercial use of massive 5GtoB applications. It includes the following key capabilities:

- Industry profiling: Builds an industry profile library based on 5GtoB projects and translates "various industry languages" into a "unified network language" to automatically match atomic service models and intelligently recommend network construction standards.
- Environment modeling: Uses the environment surveying and mapping information, such as the satellite map and point cloud, with semantic identification and other technologies to automatically complete refined environment modeling for oceans, ports, and factories, significantly improving the accuracy of network simulation planning.
- Adaptive propagation model: Builds a scenario-based propagation model baseline library based on 5GtoB projects to automatically match the most appropriate propagation model based on application scenario characteristics (sector-level adaptive propagation models can be implemented on WANs).
- User-level simulation and evaluation: Based on user distribution and service models (including toC, toH, and toB), the platform uses the Monte Carlo simulation to simulate the receive level, rate, and delay of users at different locations and time, and collect statistics on SLA compliance.
- SLA planning: Based on user-level dynamic simulation, the platform efficiently completes high-dimensional optimization on multiple indicators such as coverage, rate, and latency in batches through graph coloring, and outputs the optimal site location, RF data (including patterns), and network resource planning results that meet 5GtoB service requirements.

To meet the high SLA requirements of 5GtoB services, connection-level proactive and real-time O&M is the key. The O&M system must provide the following capabilities:

- Multi-dimensional visualization: Implement network topology and terminal location visualization, as well as network-, campus-, slice-, and connection-level performance visualization.
- Fast network- and connection-level exception detection: Use machine learning algorithms to detect exceptions in minutes based on the real-time reporting data at the service-flow level.
- Quick fault demarcation and locating: Quickly demarcate and isolate individual or mass faults based on fault modes in typical scenarios, and accurately locate root causes of network problems based on AI algorithms.
- Fault prediction and prevention: Predict device faults, significantly reducing redundant hardware backup resources and lowering network costs for industry customers.

5.4.3 Benefits

In Zhejiang, the 5GtoB intelligent O&M solutions based on the intelligent RAN architecture have been verified for the first commercial use, including E2E verification of network evaluation, network planning, visualized O&M, and intelligent fault analysis. In the public network for private use, the network coverage and rate in a specified area were accurately evaluated, greatly reducing the cost of door-to-door test and evaluation and improving the evaluation efficiency by 100x. In the industrial private network, the network site and cell resources were accurately planned, meeting the SLA requirements of different services on the live network and improving the network planning efficiency by 60%. After services were brought online, the network and terminal running status became visible in multiple dimensions, network and terminal exceptions were detected within 5 minutes, and root causes were automatically demarcated and located within 15 minutes. This helped O&M personnel quickly rectify network faults and greatly reduced O&M costs.

Intelligent and precise planning helped operators build a 5GtoB benchmark network in Ningbo Port. Based on the user track of gantry cranes and container trucks, site or RF parameters and network resources were accurately and automatically planned, meeting user requirements of concurrent port services.

6. Reference

[1] China Mobile Best Practices on Autonomous Network Whitepaper. 2021
[2] China Unicom Autonomous Network Whitepaper v2.0 2022
[3] China Telecom Autonomous Operation of Cloud Network Whitepaper.2022
[4] TM Forum Autonomous Whitepaper-time is now. 2022
[5] MTN ambition 2025.2022. [MTN 'Ambition 2025': Modern, Connected Services for Everyone | Light Reading]
[6] GTI 5G Intelligent Network Whitepaper_v1.0. 2020
[7] GTI 5G Autonomous Network Whitepaper_v2.0. 2021
[8] TM Forum. Autonomous Networks: form concept to reality. 2022
[9] TM Forum. IG1218 Autonomous Network Business Requirements and Framework. 2022
[10] TM Forum. IG1252 Autonomous Network Levels Evaluation Methodology. 2021
[11] TM Forum. IG1251 Autonomous Network Reference Architecture. 2021
[12] TM Forum. IG1253 Intent in Autonomous Networks. 2022
[13] TS28.104 Management Data Analytics (MDA)
[14] TS 28.313 Self-Organizing Networks (SON) for 5G networks
[15] 3GPP TS 28.533 "Service Based Management architecture", Release 15, 2018
[16] 3GPP TS 28.535: "Management and orchestration; Management services for communication service assurance; Requirements", Release 16, 2020
[17] 3GPP TS 28.536: "Management and orchestration; Management services for communication service assurance; Stage 2 and stage 3", Release 16, 2020
[18] 3GPP TR 28.810: "Study on concept, requirements and solutions for levels of autonomous network", Release 16, 2020
[19] 3GPP TS 28.100:"Management and orchestration; Levels of autonomous network", Release 17
[20] 3GPP TR 28.812:"Telecommunication management; Study on scenarios for Intent driven management services for mobile networks", Release 16, 2020
[21] 3GPP TS 28.312:" Management and orchestration; Intent driven management services for mobile networks", Release 16, 2020
[22] CCSA. Autonomous Network: System Architecture
[23] CCSA. Technical Requirements and Evaluation Model for Intelligent Level Evaluation of Mobile Network Management and operation
[24] CCSA. Study on intent management technologies of autonomous networks
[25] CCSA. Study on intent management technologies of mobile core networks
[26] CCSA. Technical Requirements of Knowledge management in Autonomous Network
[27] CCSA. Intelligent Level Specification for Telecommunication Network Management and Operation
[28] CCSA. Intelligent Level Specification for Telecommunication Network Management and Operation – Mobile Network
[29] ETSI. GS ENI005 System Architecture
[30] ETSI. GS ENI019 Representing, Inferring, and Proving Knowledge in ENI
[31] ETSI. GR ENI013 Intent Policy Model Gap Analysis
[32] ETSI. GR ENI015 Processing and Management of Intent Policy
[33] ETSI. GR ENI031 Construction and application of fault maintenance network knowledge graphs
[34] China Mobile. 5G-Advanced New Capability and Industry Development Whitepaper. 2022
[35] GB/T 23703.1-2009 Knowledge Management——Part I : Framework. 2009

7. Abbreviations

3GPP	3rd Generation Partnership Project
AI	Artificial Intelligence
AN	Autonomous Network
ANIMA	Autonomic Networking Integrated Model and Approach
ANL	Autonomous Network Level
CAPEX	Capital Expenditure
CI	Continuous Integration
CD	Continuous Delivery/Deployment
CT	Continuous Testing
DOU	Dataflow of usage
ENI	Experiential Networked Intelligence
E2E	End to End
ETSI	European Telecommunications Standards Institute
GTI	Global TD-LTE Initiative
KPI	Key Performance Indicator
MANO	Management and Orchestration
MIMO	Multiple-input multiple-output
ML	Machine Learning
MLOps	Machine Learning Operations
NaaS	Network as a Service
NE	Network Element
NFV	Network Function Virtualization
NMS	Network Management System
NWDAF	Network Data Analytics Function
O&M	Operation and Maintenance
OPEX	Operation Expense
PaaS	Platform as a Service
QoS	Quality of Service
SDO	Standards Development Organization
SDN	Software Defined Networking
SON	Self-Organizing Network
UE	User Equipment
XAI	Explainable Artificial Intelligence
ZSM	Zero Touch Network & Service Management