Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Position Paper:

6G autonomous radio access network empowered by artificial intelligence and network digital twin*

Guangyi LIU[†‡1,2], Juan DENG[1,2], Yanhong ZHU[1], Na LI[1,2], Boxiao HAN[2], Shoufeng WANG[3], Hua RUI[4], Jingyu WANG[5], Jianhua ZHANG[5], Ying CUI[6], Yingping CUI[1], Yang YANG[6], Yan ZHANG[7], Jiangzhou WANG[8], Ye OUYANG[3], Xiaozhou YE[3], Tao CHEN[9], Rongpeng LI[10], Yongdong ZHU[11], Yuanyuan ZHANG[9], Li YANG[4], Sen BIAN[3], Wanfei SUN[12], Qingbi ZHENG[1], Zhou TONG[1], Huimin ZHANG[1], Zecai SHAO[1], Jiajun WU[1], Mancong KANG[1]

[1]*China Mobile Research Institute, Beijing 100053, China*

[2]*ZGC Institute of Ubiquitous-X Innovation and Applications, Beijing 100080, China*

[3]*AsiaInfo Technologies (China) Inc., Beijing 100193, China*

[4]*ZTE Corporation, Shenzhen 518057, China*

[5]*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*

[6]*IoT Thrust and Research Center for Digital World with Intelligent Things, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511455, China*

[7]*Department of Informatics, University of Oslo, Oslo 0316, Norway*

[8]*School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NZ, UK*

[9]*MediaTek (Beijing) Inc., Beijing 100015, China*

[10]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China*

[11]*Zhejiang Lab, Hangzhou 311500, China*

[12]*CICT Mobile Communication Technology Co., Ltd., Beijing 100083, China*

[†]E-mail: liuguangyi@chinamobile.com

Received July 5, 2024; Revision accepted Oct. 28, 2024; Crosschecked Nov. 20, 2024; Published online Dec. 28, 2024

Abstract: The sixth-generation (6G) mobile network implements the social vision of digital twins and ubiquitous intelligence. Contrary to the fifth-generation (5G) mobile network that focuses only on communications, 6G mobile networks must natively support new capabilities such as sensing, computing, artificial intelligence (AI), big data, and security while facilitating Everything as a Service. Although 5G mobile network deployment has demonstrated that network automation and intelligence can simplify network operation and maintenance (O&M), the addition of external functionalities has resulted in low service efficiency and high operational costs. In this study, a technology framework for a 6G autonomous radio access network (RAN) is proposed to achieve a high-level network autonomy that embraces the design of native cloud, native AI, and network digital twin (NDT). First, a service-based architecture is proposed to re-architect the protocol stack of RAN, which flexibly orchestrates the services and functions on demand as well as customizes them into cloud-native services. Second, a native AI framework is structured to provide AI support for the diverse use cases of network O&M by orchestrating communications, AI models, data, and computing power demanded by AI use cases. Third, a digital twin network is developed as a virtual environment for the training, pre-validation, and tuning of AI algorithms and neural networks, avoiding possible unexpected losses of the network O&M caused by AI applications. The combination of native AI and NDT can facilitate network autonomy by building closed-loop management and optimization for RAN.

Key words: 6G; Network autonomy; Native artificial intelligence; Network digital twin; Service-based radio access network
https://doi.org/10.1631/FITEE.2400569 **CLC number:** TN929.5

‡ Corresponding author

* Project supported by the National Key Research and Development Program of China (No. 2024YFE0200600)

ORCID: Guangyi LIU, https://orcid.org/0000-0002-8656-1946

1 Introduction

The commercial deployment of fifth-generation (5G) mobile communication technology has accelerated the application of artificial intelligence (AI), cloud computing, and big data. This has promoted the integration of data technology, operational technology, information technology, and communication technology (DOICT) while improving lifestyle and accelerating digital transformation.

Several parameters and features have been specified for 5G mobile networks (3GPP, 2023c, 2023e, 2023f) to improve the performance of the radio access network (RAN). To ensure ubiquitous connection, more than four million 5G base stations (BSs) have been deployed in China and the interworking between the fourth-generation (4G) and 5G mobile communication technologies has been implemented. Optimum user experience can be guaranteed by the joint optimization of 4G and 5G mobile networks. Manual planning, configuration, optimization, and operation and maintenance (O&M) result in high complexity, low efficiency, and poor performance of networks, increasing the cost and overhead for mobile network operators. The high cost, power consumption, and complexity of network O&M are critical issues for 5G development.

Highly diverse key performance indicators are required for applying 5G mobile networks to vertical and enterprise scenarios, e.g., extremely high data rate in the uplink for quality check of industrial production, extremely low latency and high reliability for industrial control, extremely low cost for the device and private network for small business, and large connection density for logistics management. Due to the demands for additional capabilities such as AI, cloud computing and storage, positioning, and sensing besides communication connection, mobile operators have to manually bundle different products to provide a turnkey solution to vertical customers. The independence among these products causes complex interactions and management, resulting in high costs, low efficiency, and long response time for customers. Therefore, providing a cost-effective personalized and customized private network solution is another critical issue for 5G development.

These issues can be addressed using AI via network automation and intelligence. In 2016, many operators and manufacturers initiated the establishment of zero-touch network and service management for achieving "zero manual intervention" network O&M via automation. The open network automation platform (ONAP) has proposed implementing intelligent network management based on software-defined networking (SDN) and network function virtualization (NFV) (https://docs.onap.org/). 3GPP (2023b) has defined different network automation levels. In 2021, China Mobile published a white paper on AI autonomous networks, hoping to achieve L4 network autonomous driving by 2025 (China Mobile, 2021). Huawei (2023) released a white paper on autonomous networks and proposed the gradual realization of autonomous networks using network intelligence. Open RAN (O-RAN) considers introducing AI applications based on RAN cloudification to achieve intelligent network optimization. 3GPP introduced a new network element termed as network data analytics function (NWDAF) (He XW et al., 2023) in the core network specifications of Release 16, expecting to introduce AI applications for improving network O&M efficiency and service assurance capabilities. However, the majority of these studies have relied on an existing network and deployed AI in a "patched" and "add-on" manner for each case. Existing system architectures pose significant challenges in achieving the desired performance by reducing data availability, reliability, consistency, effectiveness, and efficiency and introducing new management problems.

The social development vision of sixth-generation (6G) mobile communication technology toward 2030 is "digital twin (DT), ubiquitous intelligence" (Liu GY et al., 2020a, 2020b), which depicts increasingly diverse scenarios and applications (Liu GY et al., 2022). These include holographic communication, intelligent robot assistant, digital twin human, connected senses, unmanned drones, and flying cars. International Telecommunication Union Radiocommunication Bureau (ITU-R) has published *Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond* (ITU-R, 2023). The proposed recommendations include overall objectives and key capabilities and clarify the requirements of six typical scenarios, i.e., immersive communication, ultra-massive connection, extremely high reliability and low latency, global coverage, integration of communication and sensing, as well as communication and AI. 6G mobile networks

facilitate the development of diverse and fragmented applications from different industries by realizing the transformation from a traditional mobile communication network to an advanced mobile DOICT network with natively integrated communication, sensing, computing, big data, AI, and security (Liu GY et al., 2024).

Based on 5G experience and future requirements of 6G mobile networks, the following problems need to be resolved for the deployment and operation of 6G networks: (1) How can the operational efficiency and effectiveness be effectively improved during the entire network lifecycle with increasing network deployment scale? (2) How can the deployment and operation of personalized and customized networks be flexibly and quickly supported for different scenarios by automatically tailoring network functions (NFs) and physical resources as the scenarios and future application requirements become more diverse and fragmented? (3) How can the traditional scenario-driven approaches, e.g., "patched" and "add-on," be avoided for AI deployment while minimizing the trial–error cost of AI applications to unleash the full potential of network intelligence? (4) How can the network energy consumption be reduced to achieve green and low-carbon network deployment and operation?

While recent literature (Benzaid and Taleb, 2020; Almasan et al., 2022; Hazra et al., 2024; Li LL, 2024; Yaqoob et al., 2024) has addressed individual aspects of communication, sensing, big data, computing, AI, and security, a comprehensive framework integrating all these elements is still lacking. Benzaid and Taleb (2020) believed that the zero-touch network and service management (ZSM) framework is a next-generation management system that automatically executes operational processes and tasks and that AI is a crucial enabler of fully autonomous networks. Hazra et al. (2024) discussed edge intelligence for zero-touch networks and summarized the benefits of edge intelligence. Li LL (2024) surveyed the intelligence-endogenous network architecture and technologies for 6G mobile networks. Almasan et al. (2022) introduced the concept of network digital twin (NDT). Yaqoob et al. (2024) believed that leveraging real-time data and existing technologies such as SDN and NFV along with NDTs can facilitate dynamic and fully automated network orchestration, optimization, and management. This can further promote advancements toward the autonomous, self-evolving 6G network. Niemöller et al. (2024) explained autonomous network characteristics and described intent-based service management for autonomous networks.

Herein, a system framework of high-level network autonomy for 6G RAN using native AI and NDT is proposed. This framework uses cloud-native design and NFV to achieve network self-deployment and configuration, self-optimization, self-healing, and self-evolution for solving the problems of high cost, high energy consumption, and low efficiency of 6G deployment and operation.

The rest of this paper is organized as follows: The development of network autonomy is surveyed in Section 2. In Section 3, the driving forces and requirements of 6G RAN autonomy are introduced. Section 4 discusses the basic concepts and framework of 6G RAN autonomy. Section 5 introduces the enabling technologies of 6G wireless network autonomy. Section 6 presents the prototype system verification for some RAN autonomy use cases. Section 7 discusses research challenges and open issues, and Section 8 concludes the paper.

2 Development of network autonomy

2.1 Network management

An end-to-end network comprises a wireless network, transport network, core network, data network, and other parts, as illustrated in Fig. 1. Centralized and distributed management models are used for the current operations and maintenance (Kalogiros et al., 2021). The centralized model of network O&M is typically managed by a central team or a platform that governs the configuration, operation, and maintenance of the entire network with unified decision-making and control. This model ensures network consistency and uniformity, streamlines management processes, and reduces management costs. However, it may encounter issues such as single point of failure and performance bottlenecks. In contrast, the distributed model of network O&M decentralizes management responsibilities to various regions or domains, where each domain independently manages and operates the network elements based on its own needs and conditions. This model

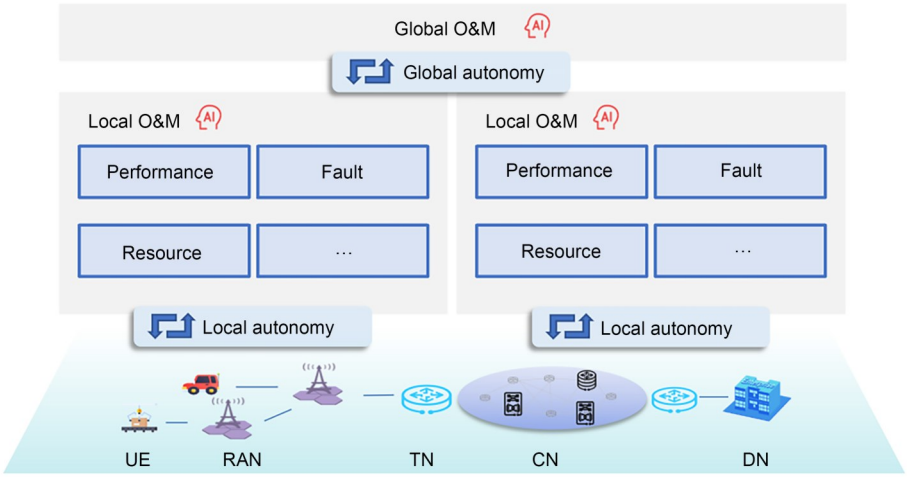


Fig. 1 Network management architecture (O&M: operation and maintenance; UE: user equipment; RAN: radio access network; TN: transport network; CN: core network; DN: data network)

can better adapt to localized and personalized requirements as well as improve the response time and operational flexibility. It may also pose challenges such as decentralized management and coordination difficulties. As networks evolve, the interconnection and interoperability between various domains become increasingly important. Therefore, the current network O&M considers centralized and distributed management models and addresses challenges related to the domain autonomy and cross-domain collaboration to achieve unified management and coordination (Ziegler et al., 2020).

Network O&M personnel must manage abundant network devices, applications, and services while addressing rapidly growing network traffic and increasingly sophisticated security threats. Many organizations are therefore adopting automation tools and platforms to simplify daily management tasks and enhance the O&M efficiency. These automation tools assist O&M personnel in automatically identifying and diagnosing network issues as well as swiftly pinpointing and resolving faults, thereby reducing manual intervention, lowering O&M costs, and enhancing network stability and reliability. Moreover, network O&M management increasingly emphasizes data-driven approaches by leveraging the monitoring, analysis, and prediction of network performance and user behavior to identify potential issues and promptly take preventive measures. Consequently, the current network O&M is evolving toward intelligence, automation, and data-driven directions to adapt to the ever-changing network environment and business requirements.

2.2 Network intelligence

As network management evolves toward intelligence, automation, and data-driven approaches, AI and advanced technologies have considerably enhanced performance and efficiency of network O&M by enabling more responsive operations and addressing modern network complexities. Recent advancements in AI have significantly enhanced network O&M by enabling automatic fault detection, real-time diagnostics, and predictive maintenance, which reduces downtime and improve service quality (Yang Y et al., 2023). Yang YQ et al. (2024) introduced TelOps, an AI-driven O&M framework that addresses the unique challenges of telecommunication networks by integrating data and empirical knowledge. Boutaba et al. (2021) proposed AI-driven closed-loop automation for 5G and beyond mobile networks using the monitoring–analyzing–planning–executing control loop to improve service orchestration and resource management. Umoga et al. (2024) explored AI-driven optimization techniques, emphasizing their potential to enhance network performance and efficiency by optimizing network configurations and resource allocations. Benzaid and Taleb (2020) discussed the ZSM framework to fully automate 5G networks as well as reduce operational costs and human errors while addressing potential AI-related risks. These studies collectively highlighted the transformative impact of AI on network O&M,

paving the way for more intelligent, efficient, and autonomous network management.

The intelligence and efficient management of O-RAN are crucial features of its open network architecture. Traditional network architectures often rely on manual configurations and management. Contrarily, O-RAN leverages its intelligence to achieve smarter network operations and employs advanced technologies such as automation, machine learning (ML), and AI to enable networks to make intelligent decisions based on real-time data and analytics. For instance, it can automatically adjust network configurations, optimize resource allocation, as well as predict and self-heal network failures, thereby enhancing the network's intelligence.

The service management and orchestration (SMO) component of O-RAN is crucial for ensuring more efficient network management. Traditional network management is fragmented across various vendors and systems; therefore, it lacks a unified management platform and standardized management processes. In contrast, the SMO of O-RAN provides a unified management interface and standardized management processes, enabling centralized management and coordination of the entire network. SMO can facilitate rapid service deployment, flexible configuration, real-time monitoring, and intelligent optimization by centrally managing various services and resources, thus improving the efficiency and quality of network management. Moreover, O-RAN's intelligence and efficient management reduce network operational costs and enhance network performance while fostering innovation and competition. Using automation and intelligence technologies, operators can more flexibly provide innovative services to meet the evolving needs of customers. A reference architecture for an O-RAN is shown in Fig. 2. As O-RAN is an open network architecture, it attracts more participants to join in and collectively drives network development and innovation. This open and innovative environment promotes technological advancements and stimulates innovation in business models, thereby advancing the entire industry.

2.3 Network autonomy

Network autonomy is the ability of a network to manage, operate, and optimize itself with minimum human intervention. Autonomous networks can dynamically adapt to changing conditions, address issues with self-diagnosis, and implement corrective actions by integrating AI and advanced technologies throughout the network stack, thereby improving efficiency and reliability. Wang S et al. (2020) discussed the transition toward 6G networks, emphasizing the need for a distributed and autonomous architecture to

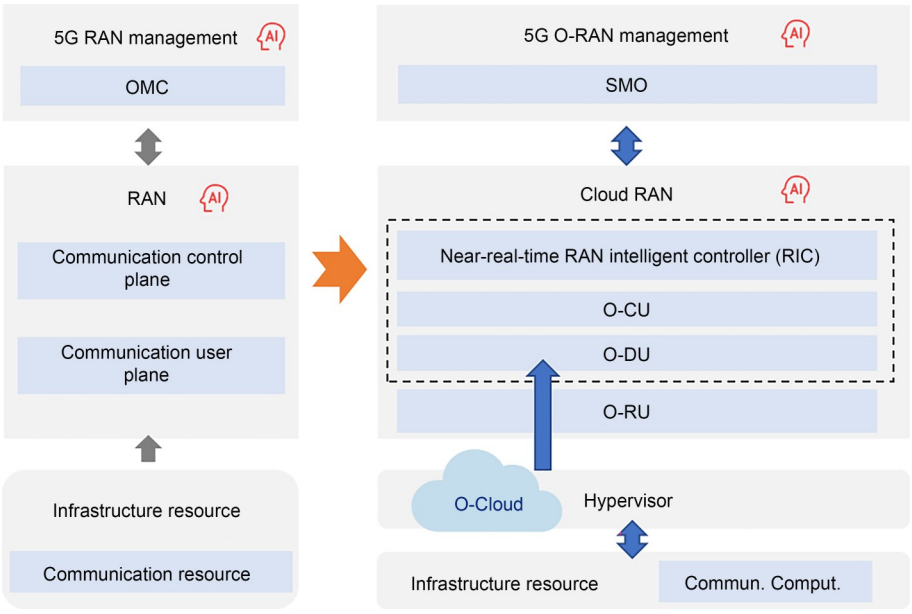


Fig. 2 A reference architecture for an O-RAN (RAN: radio access network; O-RAN: open RAN; OMC: operation and maintenance center; SMO: service management and orchestration; O-CU: O-RAN centralized unit; O-DU: O-RAN distributed unit; O-RU: O-RAN radio unit; Commun.: communication; Comput.: computation)

meet the demands of global connectivity, scalability, and self-operation. Adem et al. (2021) highlighted the critical role of AI in enabling 6G network autonomy, addressing the complexities of global connectivity, heterogeneous users, and stringent performance requirements. Mai et al. (2022) proposed a framework for end-to-end quality of service (QoS) assurance in 5G/6G networks, demonstrating how autonomous subsystems collaborate to meet QoS goals without sharing sensitive local information. Coronado et al. (2022) surveyed zero-touch management solutions for 5G and 6G networks and discussed the integration of fully autonomous network management techniques with human oversight to handle increasing network complexity and service demands.

Industry manufacturers such as Huawei are researching ways to implement network autonomy (Huawei, 2023). Autonomous driving network (ADN) leverages AI integration, deep self-learning techniques, and knowledge reasoning to autonomously interpret and address business demands within its three-layer architecture: cloud intelligence, network intelligence, and element intelligence. ADN enhances operational efficiency and reduces network failures by continuously monitoring network status, predicting risks, and autonomously rectifying issues. This approach enables automated deployment, fault prediction, event-driven self-recovery, and adaptive optimization, thereby facilitating end-to-end lifecycle management and deployment of network intents.

The ADN architecture features cloud intelligence for aggregating telecommunication knowledge and facilitates data training, model generation, and optimization. Network intelligence employs big data analytics, intelligent algorithms, and service-oriented application programming interfaces (APIs) to automate business intents, enhance network operation intelligence, and support service-oriented network businesses. Element intelligence provides real-time perception, analysis, and inference capabilities using embedded intelligent reasoning frameworks at microsecond-level precision. ADN's focus on integrating network and operational intelligence via its three-layer architecture supports mobile operators in achieving digital transformation and accelerating their path toward intelligent network management.

3 Drivers for 6G autonomous RAN

3.1 Fast response to diverse use cases

Toward the Internet of Everything, 5G mobile networks have defined three typical scenarios: enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (uRLLC), and massive machine-type communication (mMTC). Simultaneously, new features such as mobile edge computing and network slicing have been introduced to adapt to differentiated and fragmented vertical application requirements (Pivoto et al., 2023). However, vertical application scenarios often require more than traditional communication to achieve network capabilities (Nidhi et al., 2022). High-precision positioning, big data, computing, and AI capabilities can be acquired using 5G networks. However, traditional 5G devices with form factors for business-to-consumer (B2C) cannot sufficiently meet the customized needs of vertical customers for business-to-business (B2B). Vertical application demands are increasingly customized, differentiated, and fragmented. However, 6G networks enable a host of emerging use cases such as holographic communication, digital twins, and intelligent transportation systems (Kaur and Khan, 2022).

New differentiated scenarios require networks to provide new capabilities and customized services. As defined by ITU, 6G mobile networks will have new and enhanced capabilities. Specifically, the enhanced capabilities for traditional mobile communication include high peak data rate, user-experienced data rate, reliability, mobility, and connection density as well as low latency. New capabilities include ubiquitous capabilities, applicable AI-related capabilities, sustainability, interoperability, coverage, and positioning (ITU-R, 2023). 6G RAN must have capabilities beyond traditional communication networks, including communication and new capabilities, to provide on-demand services for different application scenarios and achieve Everything as a Service (XaaS) (Liu GY et al., 2024). ITU has proposed six typical scenarios for 6G mobile networks, including immersive communication, hyper-reliable and low-latency communication (HRLLC), mMTC, ubiquitous connectivity, integrated AI and communication, and integrated sensing and communication (ISAC). However, these diverse capabilities will require significant differences in network

functionality and protocol requirements. Therefore, a network that integrates multidimensional capabilities must be designed to provide different capabilities with a scalable architecture.

These use cases and capabilities present new challenges for network performance and functionality. Solution providers must therefore develop customized products tailored to the personalized needs of vertical customers, which results in issues such as long research and development cycles, high costs, low QoS, and low efficiency. From a functional perspective, a centralized unit (CU) or distributed unit (DU) with a large granularity degree is the current minimum granularity for developing a baseband unit (BBU) (3GPP, 2017). Standardization is required for updating CUs or DUs in terms of release management, with a minimum software update cycle of five years (China Mobile, 2023). However, the speed of software functionality rollout is crucial for determining the market share of a network operator; in such cases, the network cannot efficiently meet the demands of industry applications in a timely manner, which reduces the infrastructure value. Delays in feature rollout can lead to a significant loss in market shares as competitors offering more updated services gain a competitive edge. Although the release cycle can be shortened by reducing the standardization and industrialization time, network stability may be compromised due to longer downtime and service interruptions. Therefore, future network design must consider agile development, rapid introduction, and on-demand iterative evolution to meet the market demand and maintain high stability and reliability.

On-demand fulfillment is essential for 6G RAN development, for which 6G RAN must be optimized and adjusted in real time based on application demands and radio status (Shahjalal et al., 2023). It involves dynamically orchestrating functions and allocating resources to satisfy performance requirements in various scenarios. It can also efficiently allocate and use limited network resources to enhance resource utilization efficiency. To achieve on-demand fulfillment and meet future requirements for diversified QoS, management policy, deployment requirements, and openness, 6G RAN must focus on service capability. This involves decoupling and reconfiguring the functions of traditional integrated BSs into more fine-grained NFs and services. Then, the interaction and capability openness of functions are realized via service-based interfaces to provide flexible and streamlined network service during on-demand integration. To realize on-demand service provisions, (1) a resource layer must be built for use as a unified 6G network infrastructure platform for providing wireless distribution, computing, storage, and other multidimensional resources while realizing the pooling and sharing of resources; (2) on-demand orchestration of NFs and matching application requirements with the collected network and resource status is required to realize on-demand configuration of NFs and wireless resources; (3) deployment and operation modes must be able to switch seamlessly based on real-time perception of application requirements to provide multidimensional capabilities and guarantee end-to-end QoS.

Cloud-native and service-based architectures are the core technologies for realizing 6G RAN on-demand fulfillment (Li Q et al., 2022). The cloud-native architecture enables dynamic sharing and elastic scaling of underlying resources, whereas the end-to-end service-based architecture supports on-demand scheduling and integrating services, functions, and resources while enabling personalized and customized on-demand services.

3.2 O&M in advance

O&M in advance is crucial for the advancement of 6G autonomous networks and network management; it enables networks to achieve self-optimization and self-maintenance without manual intervention. The current 5G network lifecycle has separate stages for planning, construction, maintenance, and optimization, making it difficult to achieve proactive maintenance. Therefore, unified management and control based on NDTs becomes crucial in 6G mobile networks, enabling the shift from high-cost event handling to low-cost O&M in advance (Fig. 3). This transformative approach leverages advanced technologies such as automation and NDTs to create a more adaptive, resilient, and efficient network infrastructure. By continuously monitoring network performance and using AI and big data for predictive analysis, proactive maintenance will revolutionize network management. This methodology allows for the early identification

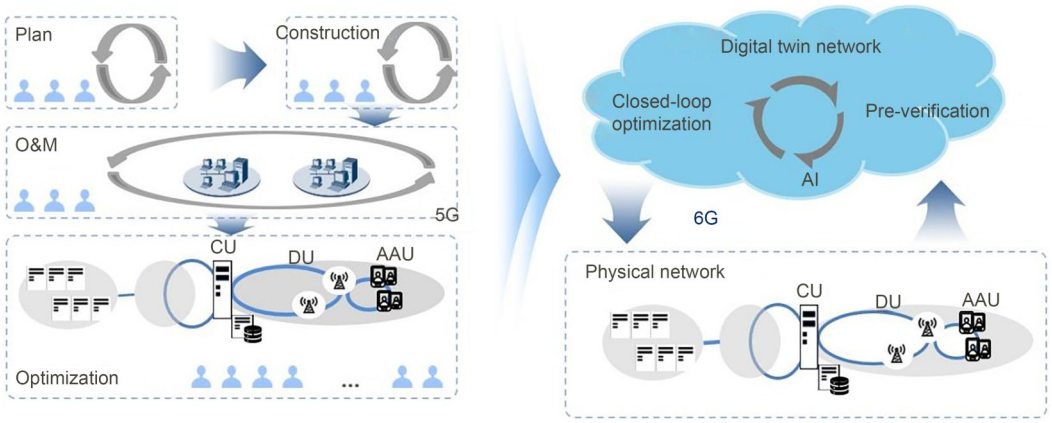


Fig. 3 Shift from high-cost event handling to low-cost O&M in advance (O&M: operation and maintenance; AI: artificial intelligence; CU: centralized unit; DU: distributed unit; AAU: active antenna unit)

of potential faults, enabling timely interventions that prevent failures, optimize resource allocation, enhance user experience, and considerably reduce operational costs (Zhang D et al., 2024).

One of the primary advantages of proactive maintenance is its capability for real-time monitoring and predictive analysis. This capability enables the network to anticipate and address potential issues before they escalate into critical problems, thus averting unexpected downtimes as well as ensuring smooth and efficient network operations (DeAlmeida et al., 2021). Proactive maintenance substantially enhances network reliability and stability by minimizing emergency repairs and replacement costs.

O&M in advance plays a crucial role in resource allocation and management optimization. Moreover, predicting user demand and network load allows for dynamic resource distribution adjustments. This proactive approach ensures that high-quality service is consistently maintained even during peak usage periods, which enhances user satisfaction and reduces operational costs. O&M in advance also accelerates the deployment and validation of new technologies. Using digital twins to simulate and test new features in a virtual environment allows for the early identification and resolution of potential issues (Cui et al., 2023). This process accelerates the rollout of new technologies while ensuring seamless integration with existing network infrastructure. Furthermore, O&M in advance considerably enhances network security by continuously monitoring for anomalies and potential threats. Real-time network behavior analysis enables quick detection and response to security breaches, thereby safeguarding the network from malicious attacks.

In summary, O&M in advance is essential for the evolution of 6G autonomous networks because it drives improvements in reliability, efficiency, and security while fostering innovation and cost-effectiveness in network management.

3.3 Intent-driven O&M

Intent-driven O&M is a modern approach to network and system management, which focuses on achieving predefined business objectives via automation and intelligent control (Banerjee et al., 2021; Yang CG et al., 2023). This methodology leverages high-level intents, which are desired outcomes or goals, to guide the configuration and operation of network systems. It captures and interprets high-level user needs and business objectives (intents to automatically generate and execute corresponding network configurations and operational strategies), thereby achieving automated, optimized, and intelligent network O&M. Intent is an abstract form of expressing the goals, needs, or expectations of a user or a system at a high level and is typically independent of underlying technical details. It reflects the expected outcomes of users or businesses regarding network services, performance, or behavior without using specific configurations or implementation methods. The "intention" should be accurately defined by stakeholders such as network operators, service providers, and standardization organizations using the weighted sum

method. Different stakeholders will assign weighting factors to various objective functions based on their respective business needs and priorities. These weighting factors reflect the relative importance of each objective in the overall optimization process. By adjusting these weighting factors, the preferences of stakeholders can be dynamically reflected during optimization under various circumstances. Thus, the "intention" can be reasonably defined and effectively implemented.

One of the critical advantages of intent-driven O&M is its ability to streamline operations. By defining intents, organizations can automate complex workflows and reduce the need for manual interventions. This automation increases efficiency and reduces human errors, thereby resulting in more reliable and consistent network performance. It enhances the agility and responsiveness of networks. Intent-driven O&M enables the network to adapt quickly to dynamic conditions and demands. This dynamic adaptability is crucial for maintaining high service quality in environments with fluctuating demands. It enhances the predictive and proactive capabilities of network management. These systems can predict potential issues and take preemptive measures using AI and ML. This predictive maintenance approach reduces downtime and improves the overall network reliability. Intent-driven O&M supports better alignment of intents with business goals. By focusing on high-level intents, network operations are directly tied to the strategic objectives of an organization. This alignment ensures that the network supports and drives a business to success. The workflow of intent-driven O&M for "zero-touch, zero-wait, zero-fault" is shown in Fig. 4, wherein its framework is divided into three main modules (zero-touch, zero-wait, and zero-fault). Each module comprises three parts, namely key characteristics, design enablers, and technologies, and interacts with the underlying physical network. The autonomous module receives intents and network data and preprocesses them to generate training data, which are then processed by the scalable intelligent intent processing module for training. Based on part of the intent data, policies are generated and sent to the decoupling module that generates instructions based on the policies and issues them to the underlying network, thereby providing services to O&M users.

In conclusion, intent-driven O&M transforms traditional network operations by introducing automation, enhancing agility, enabling predictive maintenance, and ensuring alignment with business objectives. This modern approach considerably improves the efficiency, reliability, and effectiveness of network management.

Some methods, such as data anonymization, differential privacy, and federated learning, can be adopted to achieve intent-driven O&M without infringing on the privacy of underlying pipeline content (Jain and Paul, 2013; Kim and Feamster, 2013; Hu et al., 2014). Intent-driven O&M can describe intents at an abstract level, ensuring that high-level intents can describe user and business objectives without involving specific underlying details. Alternatively, a modular

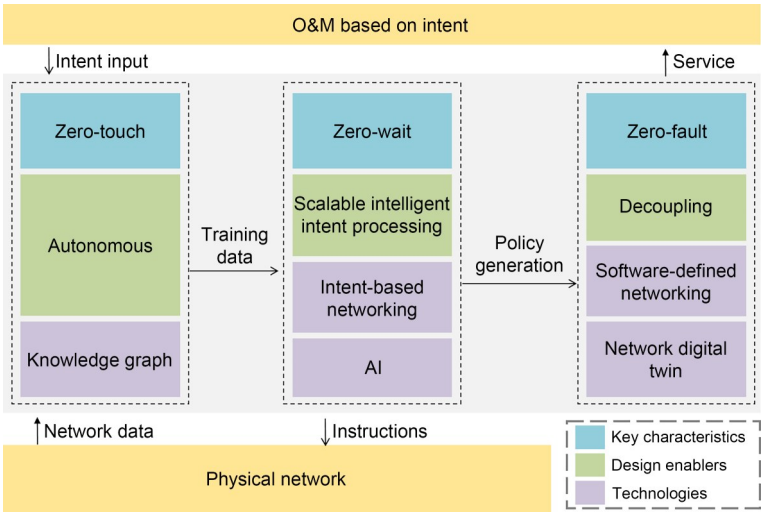


Fig. 4 Intent-driven O&M for "zero-touch, zero-wait, zero-fault" (O&M: operation and maintenance; AI: artificial intelligence)

architecture as well as a policy translation and execution engine can be used to achieve this goal.

3.4 Green method

The comparison of energy and data efficiency of 4G and 5G networks is shown in Table 1. To meet the high-throughput requirement for IMT-2020, 5G mobile networks leverage a large bandwidth (e.g., 100 MHz for sub-6 GHz and 400 MHz for mmWave), high transmitting power, and large-scale antenna technology; this considerably increases the energy consumption of 5G networks (Kamran et al., 2024). With considerably enhanced data transmission capacity of a single BS, the energy consumption for a single bit has substantially decreased; however, the increase in the absolute energy consumption has become a serious challenge for 5G network operation.

Various BS energy-saving solutions have been employed to address this issue. These include employing more advanced processes for reducing chip energy consumption, using AI technologies for automatically acquiring energy-saving strategies by predicting service loads (Mahbub and Shubair, 2022; Mao et al., 2022; Zhao et al., 2022), and deploying cell shutdown, carrier mutation (Lähdekorpi et al., 2017; Younes and Louet, 2022), radio frequency channel mutation, and symbol mutation that reduces the energy consumption of the network by approximately 15%. However, each energy-saving strategy has its specific applicable scenario and impact on network quality. Due to the limitations of existing BS designs, their baseline energy consumption at no load has exceeded 60% of the total energy consumption of the entire BS at 100% load. This has somewhat decreased the energy-saving gain of the 5G network. Overall energy consumption will continue to increase with increasing numbers of BSs and data centers, which will continuously burden operators. Therefore, more efficient and effective energy-saving measures have to be pursued in the 6G network.

4 6G autonomous RAN

4.1 Vision of 6G autonomy

As shown in Fig. 5, a 6G autonomous network must achieve "zero touch, zero wait, and zero faults" using four fully automated and intelligent processes: self-deployment and configuration, self-optimization, self-healing, and self-evolution during the entire lifecycle.

Table 1 Energy and data efficiency comparison of 4G and 5G networks

Status	Equipment model (number of channels and the overall transmit power)	Bandwidth (MHz)	Power (W) (BBU+3AAU)	Throughput (Mb/s)	Energy efficiency (J/bit)
Peak power	4G (8 and 120 W)	60	1100	120	1
	5G (16 and 240 W)	160*	3131	1216	0.278
	5G (64 and 240 W)	160*	4297	1963	0.238
Average power (30% load)	4G (8 and 120 W)	60	722	36	2.188
	5G (16 and 240 W)	160*	1798	364.8	0.538
	5G (64 and 240 W)	160*	3039	589.8	0.562
Minimum power (idle, 0 load)	4G (8 and 120 W)	60	560	0	
	5G (16 and 240 W)	160*	1226	0	
	5G (64 and 240 W)	160*	2500	0	

* 4G included. For column "Energy efficiency," the relative values to 4G mobile networks are presented. BBU: baseband unit; AAU: active antenna unit

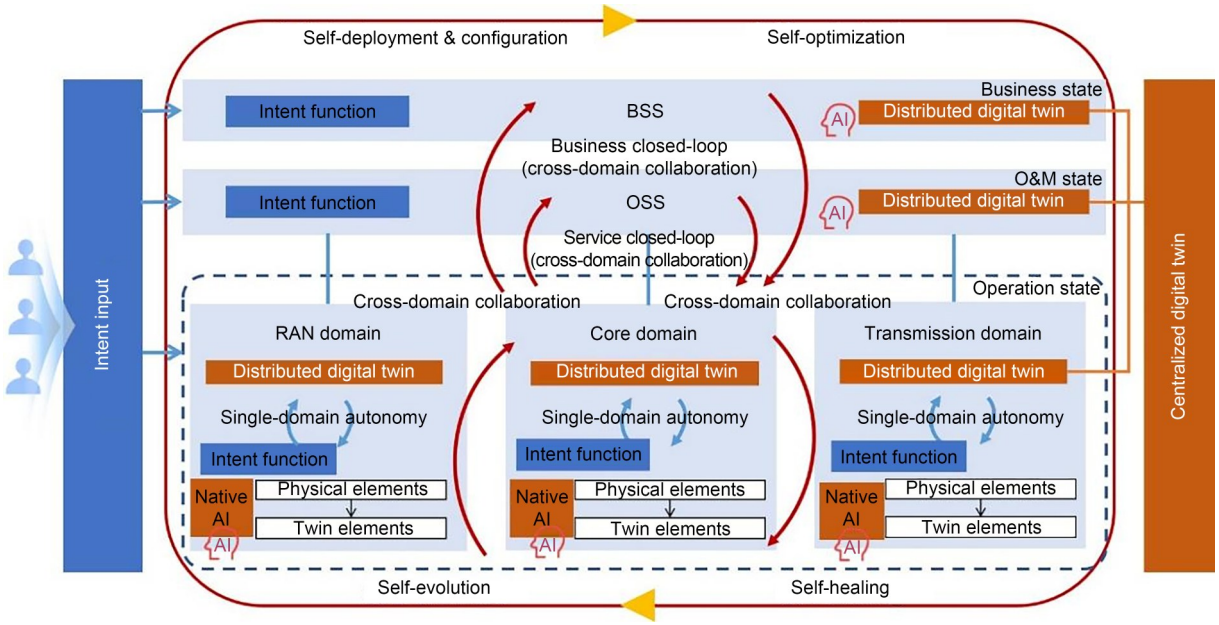


Fig. 5 Vision of 6G autonomy (BSS: business support system; OSS: operation support system; O&M: operation and maintenance; RAN: radio access network; AI: artificial intelligence)

A 6G autonomous network involves operation, O&M, and business states. Automation and intelligent orchestration and management can be achieved in each state based on intent and NDT. In the operation state, native AI and NDT enable single-domain autonomy and cross-domain collaboration.

Intelligent network elements and decision intelligence can be achieved within a single domain using native AI, thereby providing the network with dynamic policy updates and real-time adjustments. The digital twin network (DTN) offers a data generation and prevalidation environment for AI decisions, which further enhances the generalization capability of AI, prevents network failures caused by inaccurate AI decisions, and provides reliable assurance for network autonomy. In the O&M and business states, the network can automatically generate optimization and evolution strategies based on the input intent and validate them using digital twins. These strategies are then automatically converted into network configurations and implemented within a single domain of the operation state, achieving service closed-loop and business closed-loop systems.

To meet the flexible autonomy needs of a network, digital twin elements can be deployed based on a "centralized control + distributed intelligence" approach. End-to-end NDT supports end-to-end network-level intelligent requirements. Contrarily, the basic functions of digital twins within local domains, such as the core network domain, transportation domain, and wireless network domain, support a network's intelligent requirements within those domains. Moreover, native AI is deployed in a distributed manner, enabling flexible support for network element intelligence, single-domain intelligence, and cross-domain intelligence.

4.2 6G autonomous RAN architecture

Based on the above analysis, a 6G autonomous RAN architecture is proposed (Fig. 6) based on cloud-native and service-based networks, native AI, and NDT.

RAN (Zong et al., 2022) was implemented in a service-based manner based on cloud-native basic resources, with the capability of on-demand function orchestration and integration, agile deployment, and rapid iteration. It could thus meet the needs of diversified and fragmented vertical applications and quickly align with market demands. Traditional RAN protocols can be split into multiple services as well as orchestrated and integrated on demand. New capabilities such as computing, data, and AI also exist as services and NFs (Ismail and Mahmoud, 2020), helping introduce future capabilities.

A service-based AI function can perform network data analysis and decision-making, enhance network

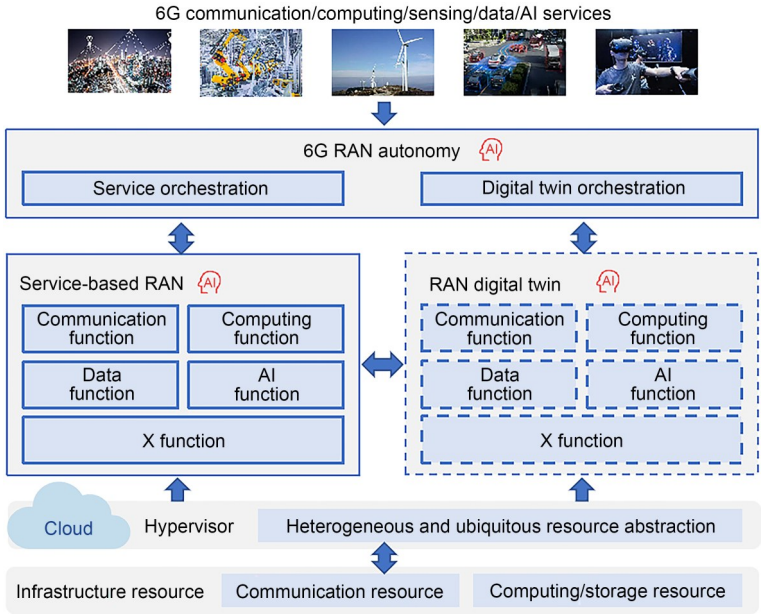


Fig. 6 6G autonomous RAN architecture (RAN: radio access network; AI: artificial intelligence)

operation efficiency, and improve network autonomy. Combined with communication, data, computing, and other functions, an efficient AI service supply system can be developed within the network to meet the needs of AI use cases within and outside the network.

For early and independent prediction of network faults and failures and achieving a high degree of autonomy with L5 (zero touch, zero wait, and zero faults), NDT represents the service-based RAN function and network environment.

With the proliferation and increasing complexity of service-oriented NFs, their effective management and coordination has become challenging. Network service orchestration is therefore required to achieve network autonomy. Network service orchestration refers to the automated and intelligent integration and coordination of various functions within a network to achieve specific service requirements or service objectives. It dynamically selects, deploys, and adjusts various functions based on user needs and conditions, thus enabling more efficient, reliable, and flexible network service delivery. A communication function is required to connect the constructed DTs and network ontology to a physical network, which allocates computing resources and operates the DTN. Therefore, the NDT requires unified scheduling to map the underlying resources and digital twin function. The orchestration of services and digital twins enables efficient and reliable network autonomy.

High-level network autonomy enhances network adaptability and self-management capabilities, thereby reducing the need for human intervention as well as enhancing the stability and reliability of a network. It also enables faster adaptation of a network to evolving environments and demands, thereby improving network efficiency and performance and providing users with a better experience and service.

4.3 Four “self-X” procedures for the lifecycle of the 6G autonomous network

The entire lifecycle of the 6G autonomous network comprises five key phases: network planning, deployment, maintenance, optimization, and operation. The proposed 6G autonomous RAN achieves comprehensive autonomous operation and management such as self-deployment and configuration, self-optimization, self-healing, and self-evolution. Self-deployment and configuration enables the network to automatically generate function combinations and parameter configurations based on the environment and demands for achieving automated deployment and configuration without human intervention. Self-optimization continuously monitors and analyzes the network status, adaptively performing global optimization of network functions and performance to

enhance user experience. Self-healing can quickly locate and fix issues when failures occur, ensuring high availability and stability of the network. Self-evolution is the ability of the network to autonomously update and upgrade in line with the technological advancements and changing demands, thereby maintaining cutting-edge service capabilities. Based on these autonomous features, the 6G network can considerably enhance management efficiency and service quality, adapt to ever-changing complex environments, and meet the diverse needs of future communications.

4.3.1 Self-deployment and configuration

Self-deployment and configuration is the foundation for achieving autonomy in 6G architectures based on service-based RAN, in which NF combinations and parameter configurations are automatically generated and delivered based on the intention and requirements without the intervention of operators. The network first creates deployment schemes for different NF combinations on-demand based on user needs and automatically determines configuration parameters according to configuration policies. These configurations are then translated into the configuration language of a physical network and delivered to it; this approach is commonly used in the planning and deployment phases of the network. In this process, the RAN digital twin that serves as an environment for generating and pre-validating deployment and configuration schemes ensures the feasibility of these configuration schemes. Various deployment plans and parameter configuration methods can be developed and tested to determine the optimal setup, which can then be applied to the RAN. This process allows for the thorough evaluation and selection of the best NF combination and configuration to ensure the best performance and efficiency of the network. The self-deployment and configuration process is shown in Fig. 7, with the detailed procedure outlined below:

1. Self-deployment and self-configuration intent is input into the RAN digital twin. This digital twin then performs self-deployment and self-configuration based on network planning and design requirements, thereby improving the efficiency of network deployment.

2. Intent parsing involves analyzing the goals and translating them into specific, actionable tasks. This step is essential for understanding the goals and forms the foundation for subsequent stages.

3. The RAN digital twin generates different NF deployment schemes by analyzing differentiated user needs. The formulations of these deployment schemes

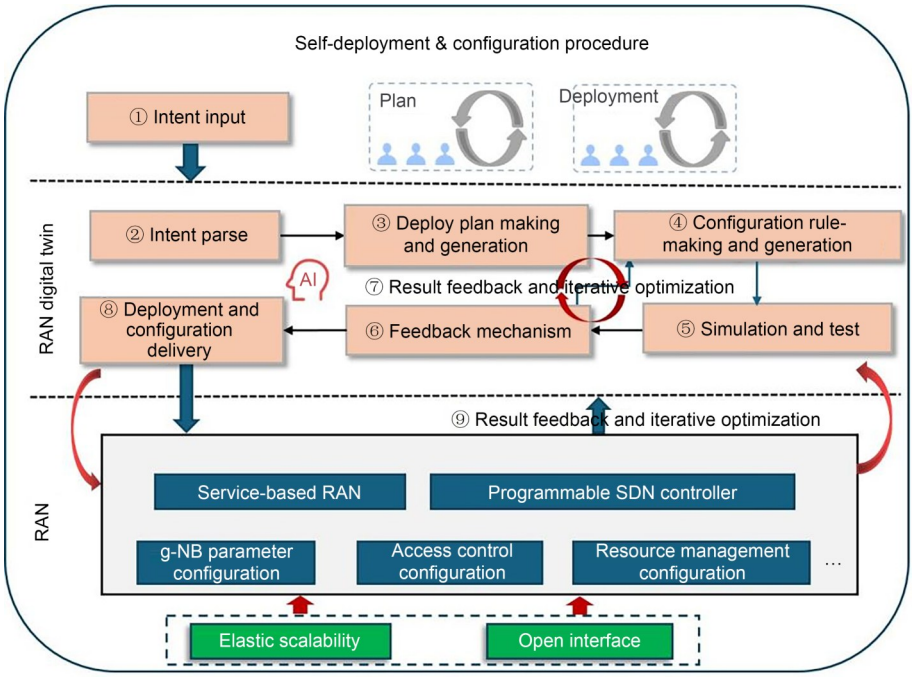


Fig. 7 Self-deployment and configuration process (RAN: radio access network; AI: artificial intelligence; SDN: software-defined networking; g-NB: next-generation Node B)

can be automated using knowledge graphs or AI-driven methods.

4. Configuration rule-making involves establishing a set of guidelines that govern the self-orchestration of NFs and resources as well as self-configuration of parameters. These rules are derived from best practices, historical data, domain-specific knowledge, and AI learning. This multifaceted approach ensures that network configurations are optimal and adaptive to changes in the network environment. By leveraging a wide range of data sources and advanced analytical techniques, the configuration process becomes more precise and robust. This ultimately enhances the network performance and reliability while reducing the need for manual intervention.

5. These rules serve as a guide for parameter configuration and settings. This involves translating guidelines into specific, actionable configurations using automated tools and scripts.

6. After deployment schemes and configurations are generated, these are subjected to simulation and test within the digital twin. This virtual space enables rigorous test without impacting the physical network. The simulation phase is critical for identifying potential issues and verifying that the function and configuration meet the intended goals. The digital twin can refine the configuration via iterative test to ensure its robustness and efficiency.

7. A feedback mechanism is integral to this process to obtain continuous insights into the configuration performance and perform real-time adjustments and fine-tuning. After validating the deployment function and configuration, they are delivered to a physical network.

8. The physical network provides feedback, which is essential for iterative optimization within the DTN. By comparing the physical network outcomes with the simulation results, the NDT can identify discrepancies for improvement. This iterative cycle of feedback and optimization ensures that the configuration evolves to meet the dynamic conditions and requirements.

4.3.2 Self-optimization

Self-optimization is an automated and intelligent process of generating a network optimization policy based on intent and requirements as well as optimizing physical network performance without the intervention of operators (Wang S et al., 2020) during the network maintenance and management phases. Using this process, the future state of the network can be accurately predicted and the corresponding optimization policies can be generated. The NDT is used to simulate and validate the optimization solutions to ensure that the policies delivered to the physical network are feasible. The validated feasible policy is then translated and automatically delivered to the physical network, thereby achieving network optimization. The self-optimization process is shown in Fig. 8, with the detailed procedure outlined below:

1. The SMO intent is input into the RAN's digital twin. The NDT can optimize the overall performance and functionality of physical networks based on the intent at irregular intervals. It can also automatically establish optimization mechanisms to periodically optimize the physical network. This flexibility ensures that the network remains efficient and responsive to varying demands and conditions. By leveraging ad-hoc and scheduled optimization processes, the digital twin entity enhances network adaptability, reliability, and performance while ensuring an optimal user experience.

2. The NDT performs real-time data sensing from the RAN and synchronizes the state of the physical network.

3. The network state prediction is then conducted. Predictive models are employed to forecast future network conditions using historical data and real-time inputs. This predictive capability is essential for proactive optimization, enabling the NDT to anticipate issues and opportunities in advance.

4. The optimization models are constructed based on the predicted state that can determine the best possible configurations and policies for a physical network. These models are designed to optimize various scenarios such as dynamic resource allocation, multi-input multi-output (MIMO) weight optimization, and the next-generation Node B (g-NB) energy-saving optimization.

5. The generated optimization policies must be pre-validated within the DTN, after which they are delivered to the physical network. If not feasible, the models are optimized and new policies are generated. If some strategies require manual intervention, such as hardware replacement, the network maintenance personnel are notified.

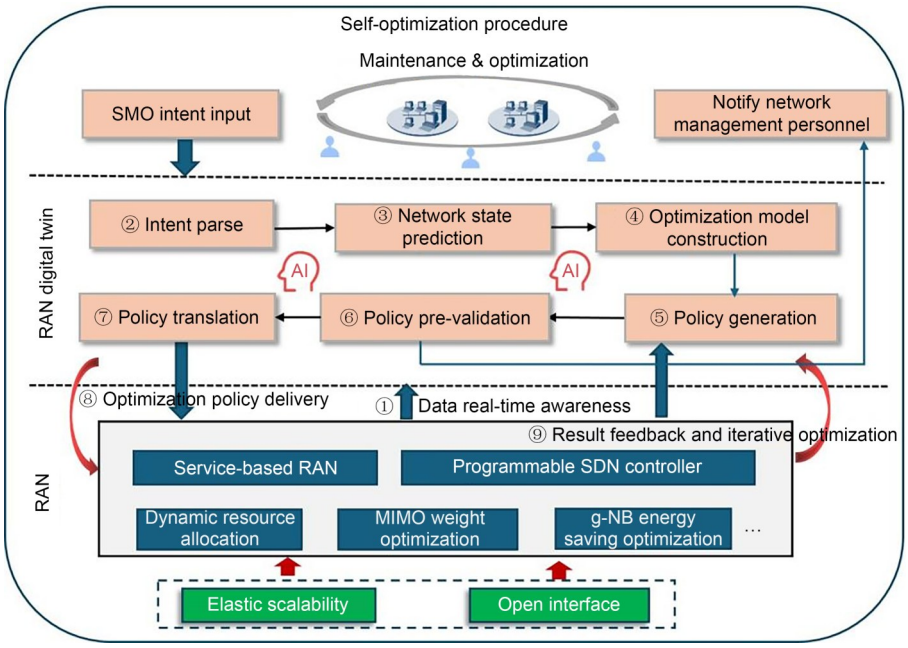


Fig. 8 Self-optimization process (SMO: service management and orchestration; RAN: radio access network; SDN: software-defined networking; MIMO: multi-input multi-output; AI: artificial intelligence; g-NB: next-generation Node B)

6. A physical network provides feedback on the optimization results, and the NDT continues to optimize the models to further improve the optimization strategies.

4.3.3 Self-healing

Self-healing is an automated and intelligent process of detecting and resolving network faults without manual intervention during the network maintenance and management phases. This process can accurately identify potential issues and implement corrective actions. NDT is used to simulate and validate healing solutions to ensure that the actions delivered to the physical network are effective.

The validated actions are then translated and automatically delivered to the RAN, thereby achieving network self-healing (Duan et al., 2022). The self-healing process is shown in Fig. 9, with the detailed procedure outlined below:

1. The self-healing intent is input into the DTN that monitors and repairs network faults based on the intent and sets up automatic fault detection and repair mechanisms. It continuously senses the network status and autonomously detects faults. This capability allows the network to quickly identify and resolve issues, ensuring high availability and reliability. By leveraging real-time data collection and advanced analytics, the NDT enhances the self-healing abilities of a network and improves its overall stability and performance.

2. The NDT performs real-time data sensing from the physical network and synchronizes its state.

3. Detection models are used to detect anomalies in network behavior using historical data and real-time inputs. This capability is essential for proactive fault detection, allowing the NDT to identify issues before they escalate.

4. The healing models are constructed based on the detected anomalies that can determine the best possible corrective policies for the network. These models are designed to address various fault scenarios such as link failures, node outages, and performance degradation.

5. The generated healing actions must be pre-validated within the DTN. After the actions are validated, they are delivered to the physical network. If not feasible, the models are optimized and new policies are generated. The network maintenance personnel are also notified if some actions require manual intervention such as hardware repair.

6. The physical network provides feedback on the healing results, and the NDT continues to optimize the models to further improve healing strategies.

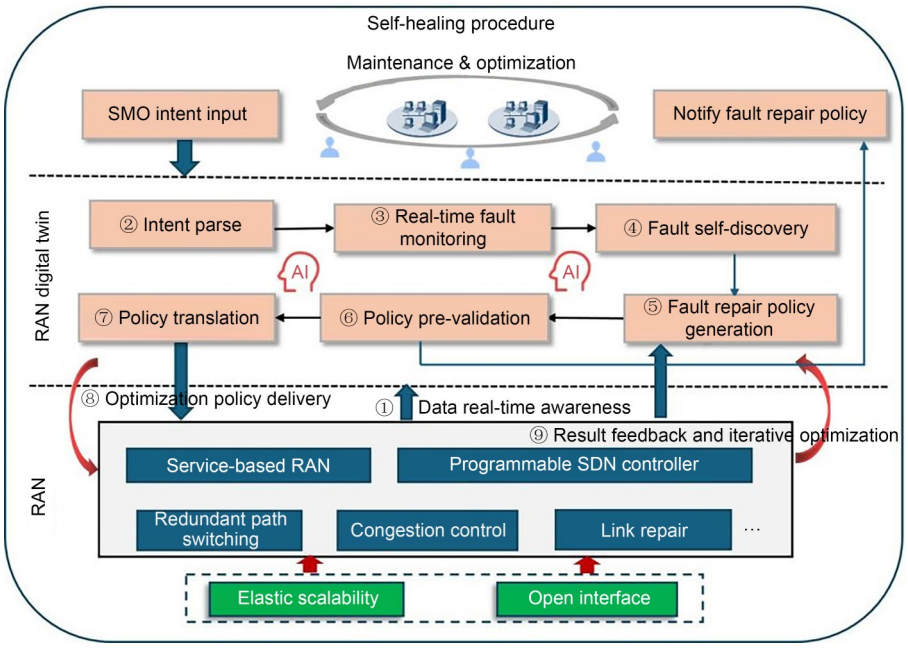


Fig. 9 Self-healing process (SMO: service management and orchestration; RAN: radio access network; SDN: software-defined networking; AI: artificial intelligence)

4.3.4 Self-evolution

Self-evolution is an automated and intelligent process that continuously discovers and upgrades new NFs, thereby enhancing network performance to adapt to new demands (Liu ZH et al., 2023). By monitoring the entire network lifecycle and via adaptive learning, it generates evolution policies, validates them, and implements function upgrades. NDT plays a crucial role in simulating, validating, and optimizing these policies before being deployed into the physical network. The physical network must support elastic expansion and open capability. The self-evolution process is shown in Fig. 10, with the detailed procedure outlined below:

1. The NDT performs real-time data sensing from the physical network and synchronizes its state.

2. The self-evolution process can be driven by intent or set up with automatic loops and evolution mechanisms using the NDT, enabling the network to self-learn autonomously even without external drivers.

3. The NDT autonomously performs dynamic monitoring of the entire network lifecycle. This comprehensive surveillance uncovers potential upgrades and enhancements to functionality and performance within the network. By continuously analyzing real-time data and historical patterns, the NDT can identify opportunities for improvements. This ensures that the network evolves to meet emerging demands and maintains the optimal performance.

4. The NDT builds network cognitive capabilities via self-learning and uses adaptive learning methods to infer and implement various network adaptive adjustment strategies to meet different user needs.

5. Evolution policies are generated to enhance network capabilities using adaptive learning methods. Large network models have powerful learning and reasoning capabilities, making them a promising adaptive learning method that can infer various network adaptive adjustment policies. Each NF unit in the DTN acts as an agent, forming collective intelligence that collaborates to generate diverse evolution strategies. Based on this collaborative effort, the network can more efficiently adapt to and respond to dynamic environments while enhancing the overall performance and reliability.

6. The DTN serves as a simulation and validation environment where various evolution policies are simulated, evaluated, and optimized. After a policy is validated as feasible, it is deployed and updated on the physical network to introduce new functions.

7. These validated evolution policies are translated into actionable steps for function upgrades. The

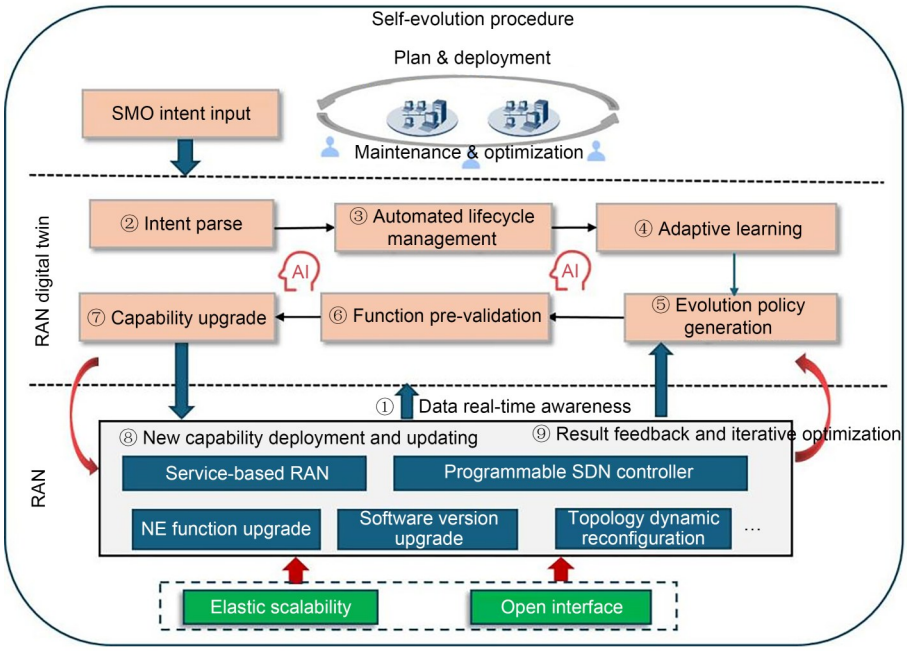


Fig. 10 Self-evolution process (SMO: service management and orchestration; RAN: radio access network; SDN: software-defined networking; AI: artificial intelligence; NE: network element)

translated policies are then deployed to a physical network, wherein existing functions are updated and new functions are deployed to meet evolving requirements. These can include software updates, protocol enhancements, and hardware improvements, which are essential for maintaining and enhancing network performance over time.

8. After deployment, the physical network provides feedback on the evolution results, and the NDT continues to optimize the models to further improve evolution strategies.

5 Enabling technologies

5.1 Cloud-native and service-based architecture

5.1.1 Concept of service-based RAN and design principles

The 6G service-based RAN based on cloud-native technologies aims to decouple traditional integrated single-BS functions into control and user plane services, achieve interaction and open capabilities between functional services using service-based interfaces, and provide more flexible and streamlined network capabilities in an on-demand combination. These functionalities help improve the network's adaptability to the entire industry (Li N et al., 2022a). This RAN includes mainly the following three technical features (China Mobile, 2022):

Feature 1: cloud user interface. Functional decoupling is a necessary step to develop a service-based RAN. By building highly cohesive and loosely coupled services, we can maximize functional reuse and concurrently leverage the advantages of cloud-native platforms.

Feature 2: service-based RAN relying on on-demand combination. By combining necessary functional services on demand, customer requirements for customization, cost control, flexible adaptation, and agile response can be met.

Feature 3: capability exposure being a basic capability provided by service-based RANs. RAN capabilities need to be exposed to more NFs or third-party applications to more effectively empower enterprise and vertical applications. Exposing directly via service-based interfaces (SBIs) can eliminate redundant peer-to-peer (P2P) interface definitions and unnecessary access and mobility management function (AMF) transparent forwarding, thereby enhancing communication efficiency.

5.1.2 Development path of service-based RAN

The evolution of service-based RANs is generally categorized into two phases: introduction of SBIs and service definition and on-demand combination (Li N et al., 2022b), as shown in Fig. 11. SBIs can be divided into partially service-based N2 interface (option 1) and fully service-based N2 interface (option 2). It can also include the openness of interfaces such as E1/F1. These two interfaces can share a phased evolution relationship, such as evolving from option 1 to option 2, or a parallel selection relationship. Option 1 is used mainly for RAN capability exposure and thus at least one RAN capability exposure service must be defined so that the core network NF (such as NWDAF) can directly invoke this service to obtain the information about RANs. Option 2 is essential for direct service invocation between the RAN NF and core network NF, and the traditional nonaccess stratum (NAS) model needs to be changed. In other words, the dependency relationship between NAS signaling must be determined.

The solution in the first stage can empower several scenarios; however, the solution in the second stage can truly realize the value of service-based architecture. In the second stage, the decoupling of RAN functionalities into services as well as the integration of RAN and core network (CN) services and procedures must be considered to better meet the design principles of high cohesion and low coupling.

5.1.3 On-demand combination

Template-based composition and dynamic composition are two viable approaches for assembling services. The former is akin to network slicing, where a template is predefined based on application needs. When an application data stream arrives, the corresponding template is invoked. The latter is well-suited for emerging application scenarios and allows selecting appropriate services as required when a data flow is encountered.

5.1.4 Service definition

1. Principles of service definition

Some relevant concepts and their relations must be first clarified, as shown in Fig. 12 (Rohani, 2023).

Function: a function is a programming structure that performs specific tasks.

Microservice: a microservice is a collection of functions that can perform target system operations and is a service component. From the perspective of continuous integration and release, a microservice is an atomic entity. Any change in the internal functionality of a microservice triggers a change in that specific microservice, and new releases are triggered through the pipeline associated with that microservice without affecting the entire system.

Service: service is a complete set of functions that are bundled together to achieve business goals. Therefore, in a broad sense, a service can be seen as

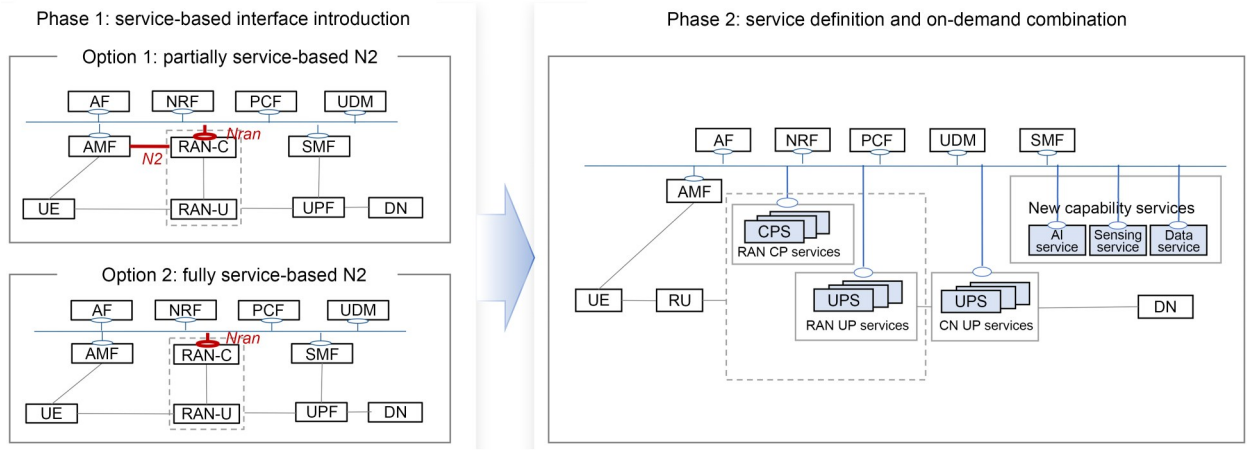


Fig. 11 Development path of service-based RAN (RAN: radio access network; AF: application function; NRF: network repository function; PCF: policy control function; UDM: unified data management; AMF: access and mobility management function; RAN-C: RAN control; SMF: session management function; UE: user equipment; RAN-U: RAN user plane; UPF: user plane function; DN: data network; CPS: cyber-physical system; RU: radio unit; CP: control plane; UPS: user plane slice; UP: user plane; AI: artificial intelligence)

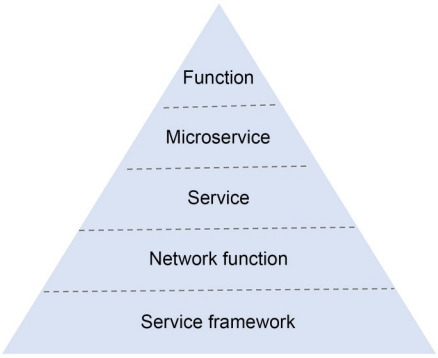


Fig. 12 Relationship among relevant concepts

a response to demand. Services are built toward business objectives and are invoked between each other through APIs.

Network function: an NF is composed of multiple services that perform related functions.

Service framework: a service framework is a universal service that can host other services, enable registration, discovery, and authorization of services, and scale and customize them for different use cases. The 5G service framework is centered around the network repository function (NRF) and service communication proxy (SCP). 6G mobile networks will further evolve with improved functions, such as enhanced NRF capabilities, to achieve automated registration, capability negotiation, state awareness, and synchronization of distributed subnets and enhance the service mesh capabilities of SCP by introducing services such as distributed routing invocation, congestion control, and disaster recovery backup.

When defining services, the following three design principles must be considered.

Principle 1: self-contained. The execution of a service should not necessarily rely on the execution of other services.

Principle 2: reusability. If a service is consumed by only one service consumer, it should not be treated as a separate service.

Principle 3: independence. The state storage involved in logical processing between different services is independent of each other.

"Self-contained," "reusability," and "independence" are three necessary principles in service design that can avoid detailed service granularity and high system complexity. When designing actual systems, the initial optimization design of services can be based on these three principles. Fig. 13 shows the process of optimizing service granularity based on the three principles of service design.

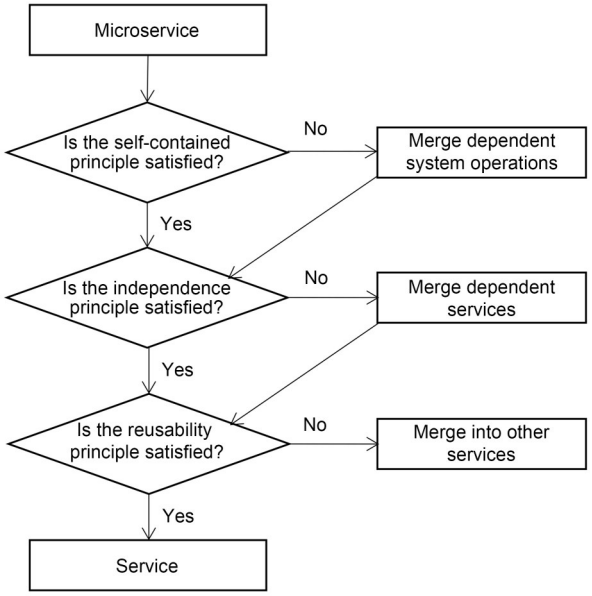


Fig. 13 Service design process

2. Definition of RAN services

3GPP specifications have a preliminary definition of access stratum services and functions. For instance, the functions supported by the radio link control (RLC) sublayer are as follows (3GPP, 2023e): (1) transfer of upper-layer protocol data units (PDUs), (2) error correction via automatic repeat request (ARQ) (only for acknowledged mode (AM) data transfer), (3) segmentation and reassembly of RLC service data units (SDUs) (only for unacknowledged mode (UM) and AM data transfer), (4) resegmentation of RLC SDU segments (only for AM data transfer), (5) duplicate detection (only for AM data transfer), (6) RLC SDU discard (only for UM and AM data transfer), (7) RLC reestablishment, and (8) protocol error detection (only for AM data transfer).

RLC entities perform the functions of an RLC sublayer. An RLC entity can be configured to perform data transfer in one of the following three modes: transparent mode (TM), UM, or AM. Consequently, it is categorized as a TM RLC entity, a UM RLC entity, or an AM RLC entity.

By combining the functions on demand, the following three services can be provided by the RLC to

the upper layer depending on the mode of data transfer that the RLC entity is configured to provide: (1) TM data transfer, (2) UM data transfer, and (3) AM data transfer (including an indication of the successful delivery of upper-layer PDUs).

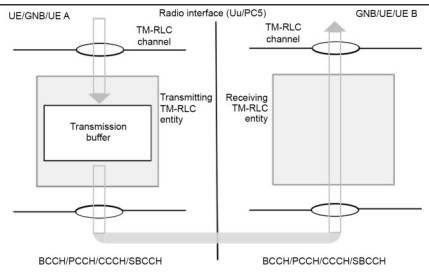
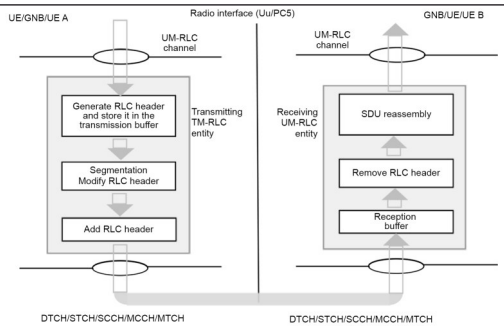
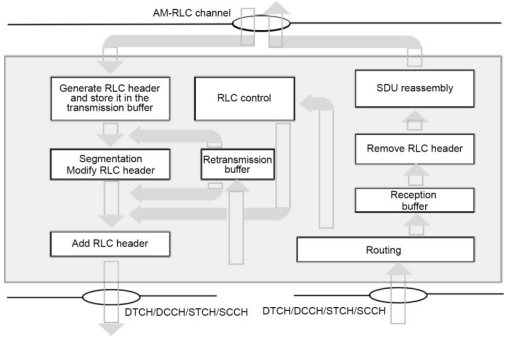
Table 2 shows the services supported by different functions. In conclusion, the following can be observed:

Not all functions can be defined as a service.

To provide a service, closely related functions have to be aggregated. For instance, RLC headers are always required after segmentation, and retransmission requests usually trigger resegmentations.

5G mobile networks were initially designed to primarily fulfill the requirements of eMBB applications. Thus, the three services provided by the RLC sublayer effectively addressed the needs of RLC at that time. As 5G mobile networks evolved, new demands and features emerged, including network coding discussed in 3GPP Rel-18. Network coding is applicable in numerous scenarios, including vehicle-to-everything (V2X), multimedia broadcast multicast service, and general downlink/uplink (DL/UL) transmissions. It uses fewer resources than packet data convergence protocol (PDCP) duplication and hybrid automatic repeat request (HARQ)/ARQ, offering higher reliable transmission. Moreover, it can be implemented at various layers. For instance, the introduction of network coding at the PDCP layer can replace PDCP duplication,

Table 2 Services provided by the radio link control (RLC) sublayer

RLC service	RLC entity	Function
Transparent mode (TM) data transfer		Transmitting a TM RLC entity: (1) do not segment the RLC service data units (SDUs); (2) do not include any RLC header in the transparent mode data (TMD) protocol data units (PDUs). Receiving a TM RLC entity: deliver TMD PDUs (which are RLC SDUs) to the upper layer.
Unacknowledged mode (UM) data transfer		Transmitting a TM RLC entity: (1) segment the RLC SDUs; (2) include relevant RLC headers in the unacknowledged mode data (UMD) PDU. Receiving a UM RLC entity: (1) detect the loss of RLC SDU segments at lower layers; (2) reassemble RLC SDUs from the received UMD PDUs and deliver the RLC SDUs to upper layers as soon as they are available; (3) discard received UMD PDUs that cannot be reassembled into an RLC SDU due to loss at lower layers of a UMD PDU which belongs to the particular RLC SDU.
Acknowledged mode (AM) data transfer		AM RLC entity Transmitting side: (1) segment the RLC SDUs; (2) include relevant RLC headers in the UMD PDU; (3) retransmit RLC SDUs or RLC SDU segments (ARQ). Receiving side: (1) detect whether or not the acknowledged mode data (AMD) PDUs have been received in duplication and discard duplicated AMD PDUs; (2) detect the loss of AMD PDUs at lower layers and request retransmissions to their peer AM RLC entity; (3) reassemble RLC SDUs from the received AMD PDUs and deliver the RLC SDUs to the upper layer as soon as they are available.

whereas it ensures ordered data delivery and reduces data caching needs at the RLC or medium access control (MAC) layer. To effectively support the new network coding features, current RLC services require further enhancements such as defining a service and the corresponding RLC entities that can accommodate network coding requirements. Alternatively, segmentation/reassembly can be established as a standalone service (similar to the PDCP encryption/decryption service), and encoding services can be integrated for fulfilling these new network requirements using tailored service combinations.

Fig. 14 shows 3GPP's definition of "service." Collectively, these "services" adhere to the three fundamental design principles of "self-contained," "reusability," and "independence." However, these "services" are conceived within the open system interconnection layered protocol model, each representing the services provided at a distinct layer. A holistic service-based architecture might overcome the inter-layer invocation limitations inherent in the layered 6G protocol model. Layers including service data adaptation protocol (SDAP), PDCP, RLC, and MAC might become obsolete and all services will constitute a singular service pool, thereby enabling flexible invocation and on-demand composition. In response to emerging requirements, additional services such as segmented services may have to be defined.

Overall, RAN user plane services comprise function-related services, such as header compression,

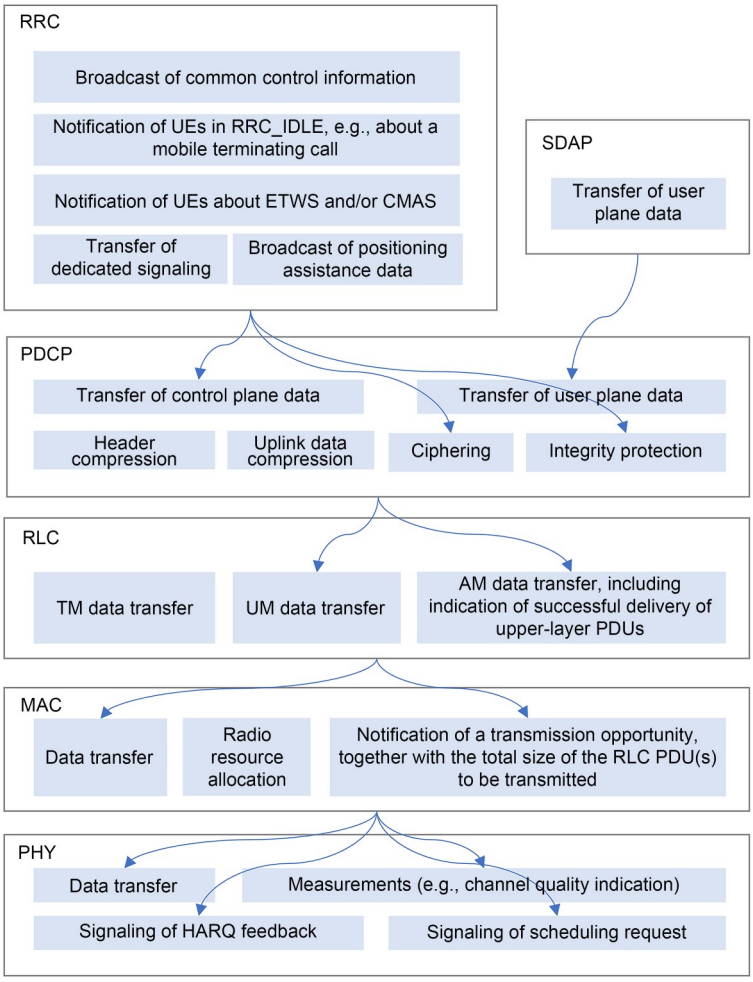


Fig. 14 Service defined by 3GPP (3GPP, 2023a, 2023c, 2023d, 2023e). RRC: radio resource control; UE: user equipment; ETWS: emergency telecommunications warning system; CMAS: commercial mobile alert system; IDLE: idle mode; SDAP: service data adaptation protocol; PDCP: packet data convergence protocol; RLC: radio link control; TM: transparent mode; UM: unacknowledged mode; AM: acknowledged mode; PDU: protocol data unit; MAC: medium access control; PHY: physical layer; HARQ: hybrid automatic repeat request

and data transmission services tailored to specific needs. The latter services are provided externally by packaging multiple tightly coupled functions together such as AM data transmission services.

5.1.5 Use case of service-based RAN

A service-based RAN introduces new nodes, functions, and features. It transcends the traditional layered protocol model, thereby enhancing the flexibility of service composition. This flexibility manifests in the number of services that can be assembled as needed and in the invocation relationships between services.

By executing multiple services in parallel, processing efficiency can be considerably enhanced (Fig. 15). For instance, the current PDCP layer performs integrity protection, followed by encryption and addition of PDCP headers. The encryption primarily targets Internet protocol (IP) and its upper-layer packet headers as well as data fields. Since MAC-I encryption is less critical, services such as encryption and integrity protection can be executed concurrently. This will ultimately enhance the data processing efficiency.

RAN resource utilization can be effectively improved via dynamic scaling at the service level. Tidal phenomena are common in networks, with temporal fluctuations in user connections and data transmission. Based on observed network tidal changes, service-based RANs can dynamically deploy or remove service instances and achieve dynamic scaling at the service level. When the number of connected users increases, only the services related to user management in the control plane are expanded. Conversely, when the downlink data volume increases, more downlink service instances are simply added. This approach differs from existing 5G data radio bearers (DRBs) that support only bearer-level scalability (Cha et al., 2022; Choi et al., 2022).

5.2 AI: evolution from case-driven and add-on to native

5.2.1 Design principle of native AI

The ITU IMT-2030 framework report identified the integration of AI and mobile networks as one of the six key scenarios for 6G mobile networks. AI is crucial for meeting new mobile communication network indicators and achieving high-level network autonomy. AI-enabled networks can excellently improve network operation efficiency, reduce O&M costs, and enhance user experience (Jiang W et al., 2021). Network autonomy optimization decisions are presently commonly realized via offline training and preinstalled AI models (Bonati, 2022). However, this approach fails to ensure real-time data effectiveness and consistency, resulting in the suboptimal performance

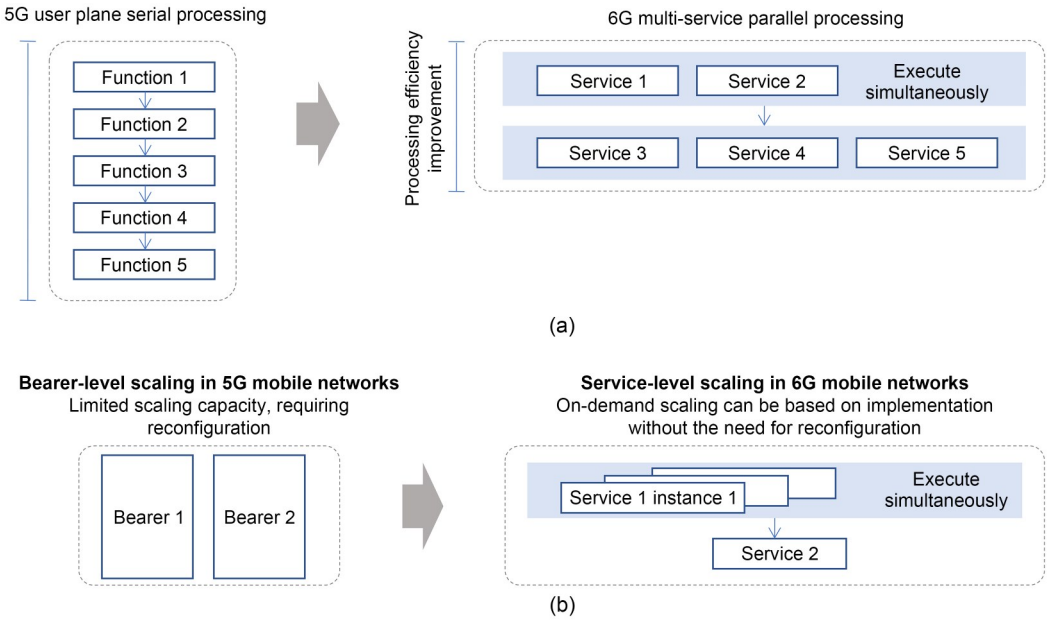


Fig. 15 Use case of service-based radio access network (RAN): (a) transformation 1 (serial → parallel processing); (b) transformation 2 (bearer-level scaling → service-level scaling)

of AI models. Moreover, implementing the entire AI process is difficult, including data collection, training, inference, optimization, and validation within the existing network infrastructure; this may lead to an unacceptable cost of trial and error. Network autonomy optimization typically involves creating individual AI models for specific use cases such as performance prediction and compression, as well as network optimization use cases such as load balancing, mobility enhancement, and beamforming. This fragmented approach hampers the rapid deployment of AI services tailored to diverse scenarios. Consequently, the current plugin and fragmented network intelligence solutions and cloud AI service supply schemes exhibit inefficiencies and fail to provide near-real-time, high-performance AI applications and services. Moreover, they do not meet the demand for intelligence in future networks with high levels of autonomy.

6G mobile networks must be therefore innovatively designed to support native AI and address these challenges in terms of network architecture and key technologies. For this, paradigms from information technology and big data must be deeply integrated with mobile communication technology. The network design will undergo four major changes, accompanied by the corresponding design principles.

1. From plugin and scenario-driven AI to native and capability-driven AI. Incremental AI function development in existing 5G architectures results in rigid structures that struggle to provide flexible and efficient AI services. 6G mobile networks must support differentiated AI from the outset through the native design. Native AI capabilities must be embedded within network's service processes and a service-oriented approach must be adopted to achieve flexibly combined NFs, ensuring unified and efficient orchestration.

2. From siloed multielement to coordinated multielement. The current approach in 5G mobile edge computing or cloud AI, where communication, computing, data, and intelligence resources operate independently and in a nonreal-time manner, lacks necessary coordination required for achieving optimal performance. To improve the efficiency of AI, 6G mobile networks should aim for globally unified, on-demand scheduling of these resources by adopting a task-centric approach in the architecture. Therein, control, communication, computing, data, and model resources are integrated to provide on-demand support for AI tasks such as inference and training.

3. From best-effort to QoS assurance. Present AI solutions lack real-time perception of user requirements, resulting in best-effort services. In response, 6G mobile networks must ensure low-latency and high-reliability AI as a service. Therefore, in terms of architecture design, unified AI service quality evaluation metrics and an end-to-end guarantee mechanism must be designed for AI service quality.

4. From high-cost postprocessing to low-cost preintervention. Currently, the effectiveness of network AI models can only be verified retrospectively, which creates a conflict between the probabilistic nature of AI and the reliability requirements of a network. To address this, 6G mobile networks must enable pre-verification of AI model effectiveness, online evaluation, and fully automated closed-loop rapid optimization. For architectural design, a native intelligence framework closely integrated with NDTs must be designed and low-cost virtual environments must be constructed to predict future network states and validate AI decisions.

In conclusion, 6G mobile networks can transform its network architecture to support native AI by incorporating these design principles. This will ensure efficient, flexible, and high-performance AI services that meet the evolving demands of future networks.

5.2.2 Native AI framework for network autonomy

Based on the aforementioned design principles and simplified intelligent network architecture, a cloud-based, service-oriented, and layered control framework is proposed herein for 6G networks with native AI capabilities. Native AI involves constructing a high-efficiency, high-performance AI service provisioning system embedded within the network architecture (Wu et al., 2021). Data, AI models, and computing resources are embedded within the network as fundamental resources, alongside connectivity. This integration allows them to be accessed and orchestrated in real time at the same architectural level, providing equal priority. By positioning these resources on par with connectivity, the network can dynamically allocate them based on specific application needs. This approach allows for task-centric, unified, and

flexible orchestration and scheduling control tailored to various scenario requirements. NDTs are also used to enhance AI reliability. The network can flexibly and efficiently provide high-quality AI services using a unified architecture. As shown in Fig. 16, the native AI framework introduces new functions in the management and orchestration platform, control plane, and user plane by considering factors such as diverse real-time demands at various stages of AI task management and the scope of control.

Within the management and orchestration platform, an AI-based network orchestration function is introduced to handle the generation and import of AI service requirements as well as the orchestrations of AI service function chains. It efficiently deploys AI NFs near the service access point, tailored to the specific service types that have been demanded. Using a series of tunneling and traffic redirection strategies between network services, it schedules and connects network services in a particular order to form a complete service chain. Consequently, it meets the service layer's QoS, i.e., QoS of AI service (QoAIS). This is typically a nonreal-time process.

The control plane incorporates AI task management, communication, computing, data, and AI control functions. The AI task management function performs native task lifecycle management and task control based on the entire lifecycle processes of AI services, including data collection, processing, training, aggregation, deployment, and monitoring. This approach aims to strike a balance between task scope and real-time scheduling, thereby ensuring the satisfaction of task-level QoS, i.e., task QoS. Further, it disaggregates this QoS into a resource QoS, which is then sent to communication, computing, digital, and intelligent control functions. These control functions enable on-demand management, including setting up and configuring communication bearers, executing computing tasks, and managing data execution functionalities. This implementation is expected to ensure the on-demand selection of AI models and successfully meet the resource-level QoS, i.e., resource QoS.

The task execution function in the user plane executes specific tasks and applies possible business logic on information and data interaction. This includes communication bearers, computing task execution, and data execution functions. A single service request may be mapped or decomposed into multiple tasks that are executed by various task execution functions. During task execution, different execution functions may interact such as the exchange of intermediate gradient information between nodes during federated learning (Niknam et al., 2020). Each task execution function can handle a single task or support multiple parallel tasks such as computation, data processing, AI training, and AI inference.

By orchestrating AI service function chains, centrally managing and scheduling AI tasks, and ensuring hierarchical support and feedback for communication, computing, data, and AI functionalities, the network can guarantee the QoAIS. This comprehensive approach ensures that 6G networks can effectively

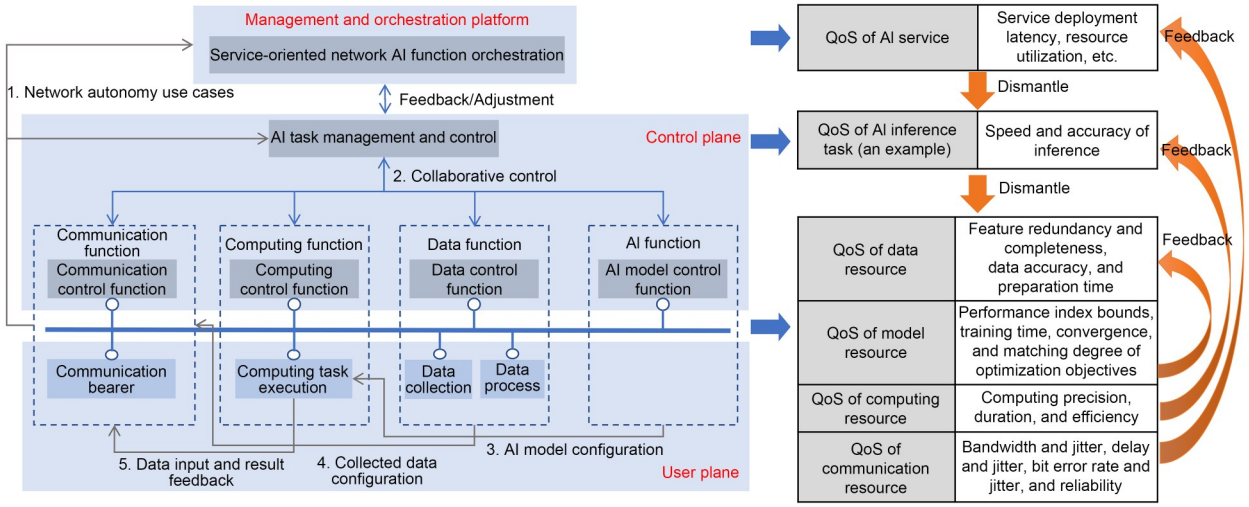


Fig. 16 6G native AI network framework (AI: artificial intelligence; QoS: quality of service)

meet the high standards of intelligence and autonomy expected in future mobile communications.

5.2.3 Data collection and management

Data are crucial for modern AI and ML applications that require massive, high-quality training and test data. Incomplete or erroneous datasets may yield unreliable models, ultimately resulting in poor decisions. However, high-quality textual data will be unavailable by 2026 (Villalobos et al., 2024), which poses a considerable threat to improving AI and ML models. Moreover, data collection is a complex process and consumes tremendous manual work and time. These problems pose a critical challenge to the access of sufficient high-quality datasets for AI and ML models.

AI data contain information required for constructing and training AI models for inference. Using inherently distributed data and computing resources, 6G network with a built-in AI capability will leverage the data source and data fabric for AI data storage and transportation.

Network data are generated/consumed by the DTN. Large amounts of network data that reflect the physical network states are required for training AI models to generate a virtual twin network. However, the transportation mechanism of the network data, which is mainly within the management domain of today's network, does not apply to the DTN.

As the components of a typical telecom network architecture, the control plane and user plane have been widely adopted in current mobile networks to provide wireless access services to mobile users (Long et al., 2022). One of the functionalities of the control plane is to build a connection and control how the user plane data are forwarded; this process is also known as session management in 5G core. User plane focuses on forwarding network packets to the right destination.

6G network is expanding its single capability, i.e., providing connectivity services for user equipment (UE), for providing AI and sensing services for any user with access to them. The network design principle (paradigm) is shifting from connection-centric to data-centric (Yan et al., 2021). These two planes are thus not applicable to data-driven services for the following reasons:

Data derived from 6G networks are abundant and expensive for the current control plane, e.g., an SBI that features reliable transport of small data packets and short-term connections.

Data topologies of emerging 6G services have become complicated with diverse producers and consumers, including RAN nodes, UE, and NFs, which are no longer connected in a simple point-to-point manner. Data aggregation, fusion, and distribution are commonly used in data topologies. However, in the current user plane, e.g., protocol data unit (PDU) session, only a one-to-one communication tunnel is set up between the UE and user plane function (UPF); this setting is inappropriate for 6G services.

Data should be exposed to execute data management or data market for trading. However, the existing PDU session makes the data flow invisible to BSs, hindering data circulation.

6G services have new requirements on subscription management for the AI service subscribers, as well as different policy controls on QoS, charging, security, and sustainability per data pipe.

To this end, a novel architecture for data management and processing must be designed to enhance data efficiency for AI and ML models.

Inspired by the separation of control plane and user plane in the CN, an independent data plane is introduced into 6G network. This will improve the flexibility and efficiency of the overall network system and provide different data process functions to collaboratively deliver the outputs required by the business (Qin et al., 2023). The data plane has the following advantages: (1) It is data-centric and focuses on architectures, and it enables data flow with any topology for flexible data services and on-path data processing. It also decouples the entire business requirement into several data process functions, including data collection, preprocessing, and analytics, and coordinates several network components responsible for efficient data transmission rather than just control signaling. Thus, it provides vast network capability for storing high volume of data. (2) The data plane enables efficient data processing at various network elements by supporting one-to-many data consumption models and flexible topologies. (3) It allows data to be exposed on collecting nodes while achieving better data security and privacy mechanisms. (4) The data plane introduces optimized data paths that reduce processing time for AI/ML applications, thereby improving the

performance and user experience. It also supports the creation of customized data topologies, allowing UE to access more personalized and adaptable services. Overall, transition to 6G mobile networks necessitates a shift toward a data-centric architecture, and the data plane can provide the necessary features to meet emerging data challenges. Three typical techniques in the data plane have been subsequently discussed.

Existing solutions are deficient for cleaning and filtering large amounts of data at the data source (Maharana et al., 2022), resulting in poor data quality that affects subsequent analysis and applications. Moreover, the data collection and processing workflow is not unified, which leads to missed or duplicated data collection. Satisfying the data collection and processing requirements of typical 6G AI services is difficult, which may impact the performance and functionality of 6G systems. Therefore, these shortcomings must be addressed by developing a comprehensive plan.

A new policy-based data collection strategy known as match and action (M&A) is proposed herein. M&A builds a unified data collection mechanism based on the data plane to avoid missed or repeated collection, which adapts to the cloudification trend with flexible deployment and good scalability. Specifically, data operation/data control (DO/DC) allocates match tables and action sets to nodes equipped with data agents (DAs) and allows the data source to execute data ingestion. The raw ingested data can be preprocessed, filtered, relayed, and analyzed with respect to file, event, network element (NE), and flow by the DAs based on the action sets defined during data transmission. In this regard, M&A pronounces an advantage in data collection, filtering, and processing for 6G data services. M&A simultaneously achieves data processing and transmission before arriving at the destination, which is beneficial for reducing the data transmission bandwidth. When the UE acts as a data provider and collects data for the DTN, it is advantageous to decouple data collection and reporting at time intervals, providing greater flexibility. This allows UE to transmit data during off-peak time points to avoid network congestion and manage power consumption more efficiently. M&A improves resource efficiency by allowing data to be logged without immediately consuming the uplink bandwidth; this is particularly helpful in areas with weak network signals or when the UE is conserving energy. Thus, the proposed approach prolongs UE battery life and contributes to a more stable network performance.

To simplify the connect-then-communicate procedure, tackle the dominating AI data, support the one-to-many data consumption model, and support efficient data distribution, a new logical function termed data commutation protocol (DCP) is proposed. DCP comprises an adaptation layer and an enhanced distributed message queue. Unlike previous architectures, wherein data are ingested into the NF via UPF, DCP serves as a middle data collector to first receive ingested data and then distribute them to the corresponding NF. By decoupling producers and consumers, DCP enables efficient asynchronous data exchange as producers do not have to wait for the response but can leave until the response arrives. Moreover, it has significant scalability with respect to supporting different client transmission protocols, e.g., transmission control protocol, user datagram protocol (UDP), and quick UDP Internet connections. This reduces producers' requirements and improves communication flexibility. Meanwhile, DCP supports many types of producers and consumers, including UPF, service data plane function, network exposure function (NEF), RAN, application function (AF), data storage function, and AI models. It also enables simultaneous real-time stream data processing and peak-clipping processing, which pronounces an improvement compared with current SBI, DU session, message queuing telemetry transport, and Kafka. With good flexibility, DCP can be deployed as an independent NF or colocated with existing NFs. Fig. 17 shows an example of DCP in an AI task: multi-task face recognition. Without a complicated connection, DCP builds a bridge from the sensing data to multiple hybrid face recognition models, including human face, age/gender, and dog breed detection models. The data collection and model inference processes are asynchronous, which enables collecting additional sensing data while the model performs predictions.

Efficient and secure data storage is crucial for AI applications, as the collection and storage of data (including training data and trained models) are critical to their performance and security. In the context of 6G networks, various AI tasks will be developed

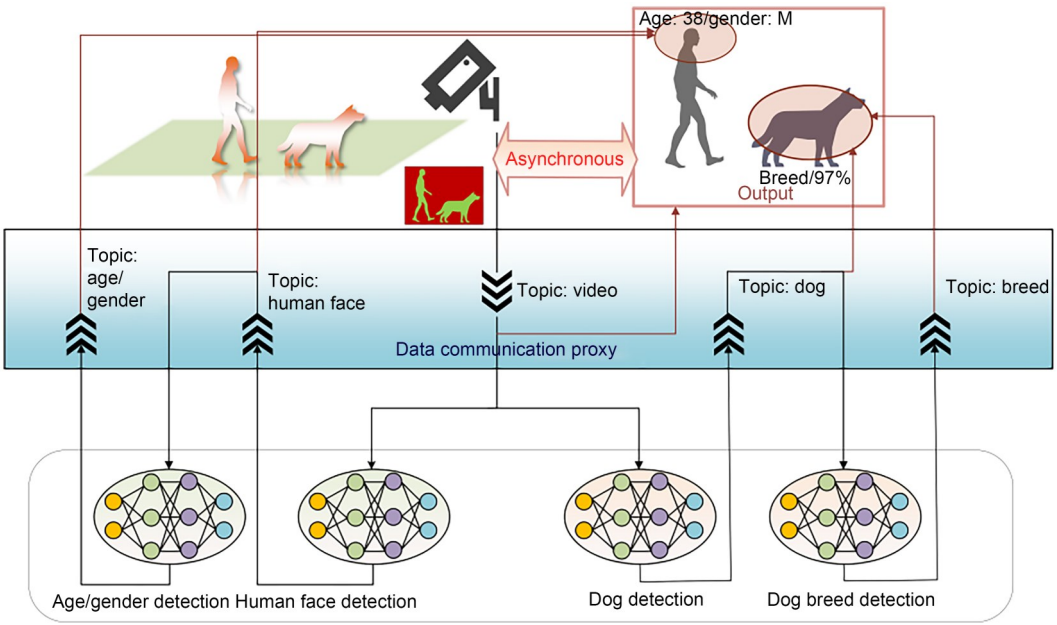


Fig. 17 An example of DCP in an AI task: multi-task face recognition (DCP: data commutation protocol; AI: artificial intelligence)

based on network data that often contain hidden patterns reflecting network or subscriber behavior. Each task requires different types of data, necessitating standardized data collection processes and the provision of appropriate data services to consumers. Robust data access controls must be implemented to allow only authenticated and authorized users to access the data and ensure security. Additionally, distributed data storage solutions should be employed to mitigate the risk of single points of failure during attacks or node failures, thereby enhancing the resilience and reliability of the system.

5.2.4 AI algorithms for network autonomy

As AI tasks on the network have become increasingly complex, the number of parameters, data volume, and computing resources required by AI models have dramatically increased (Gill et al., 2022). The capabilities of a single network node are usually insufficient for training such models, which has increased the demand for distributed learning algorithms.

Different from information technology (IT) computing clusters, a mobile network is an integrated system comprising distributed devices from multiple vendors. The distributed learning algorithms on 6G network must support various collaborative scenarios across different devices, technology domains, and heterogeneous models. The key characteristics of mobile networks, which lead to key issues that must be addressed when applying distributed learning algorithms, are described as follows.

Key issue 1: how to deal with the hyper-heterogeneity of networks?

Physical devices and terminals in the network are supplied by different vendors, resulting in complex network heterogeneities such as heterogeneous communication links, heterogeneous computational power of nodes, heterogeneous data among nodes, and heterogeneous models. These heterogeneities pose challenges for efficient distributed learning. To address the hyper-heterogeneity of networks, some effective technologies have been proposed in recent years such as hierarchical clustering learning, personalized federated learning, model optimization with knowledge distillation, and model parameter increment reporting aggregate.

Key issue 2: how to deal with the hyper-discreteness of networks?

NEs, BSs, and UE on the network are distributed in different areas and are physically separated. BSs are not connected via direct connection channels and may be composed of complex bearer networks. UE is connected to the network via an air interface, with highly valuable bandwidth resources. Therefore, when

a distributed learning algorithm is applied to a network, the communication overhead has to be further reduced and the learning efficiency has to be improved. To this end, the impact of communication process on the performance of distributed learning algorithms must be analyzed and system design in wireless networks must be guided, such as model compression and algorithm optimization.

Key issue 3: how to deal with the hyper-dynamic nature of networks?

Building an AI execution environment in wireless networks is a dynamic process. This is because the wireless connection between the terminal and the BS is affected by surrounding environments and changes in the operational status of other BSs. These changes can affect the uplink and downlink connection rates between a terminal and a BS. Moreover, the terminal has high mobility and may move from the cell center to the edge, handover to other cells, or even leave the wireless coverage area. Unlike dedicated computing resources in IT clouds, wireless networks undertake traditional mobile connection tasks and the mobile traffic carried by each network element changes constantly. The BS switches between busy and idle periods, and the terminal enters an idle mode in the absence of uplink or downlink traffic. These changes provide mobile wireless networks with the characteristics of ultra-dynamicity, which can negatively affect distributed learning and decrease model performance or result in learning failure.

Key issue 4: how to deal with the hyper-scale of networks?

A large number of nodes can participate in distributed learning in wireless networks. This distributed collaboration can break geographical limitations and enable cross-regional cooperation. For instance, distributed learning can be deployed between devices in different provinces or cities. To organize large-scale distributed learning in the network, sites and networks must excessively coordinate; this remains a significant challenge for ensuring high collaboration efficiency and performance. In some distributed learning algorithms, a large collaboration set can exponentially increase the solution space complexity, making the optimization problem unsolvable. To mitigate these challenges, semantic communication can be employed to compress information. This will considerably reduce the signaling overhead by transmitting only semantically meaningful data that are particularly beneficial in hyper-scale networks. Furthermore, edge learning enhances the ability of edge devices to perform local learning, thereby minimizing the need for backhaul communication and improving real-time processing capabilities. Wireless sensing can empower edge devices with advanced sensing abilities and enhance the accuracy of environment information retrieval, which is crucial for improving situational awareness. Moreover, AI techniques such as deep learning, reinforcement learning, and large language models (LLMs) can be applied to optimize and manage distributed learning processes, thereby reducing the dimensionality of solution spaces and addressing scalability concerns. These approaches enhance the efficiency of large-scale distributed learning and align with the integration of intelligent solutions required for handling the complexities of hyper-scale 6G networks, as pointed out by recent studies in edge learning and distributed signal processing.

Key issue 5: how to deal with the resource management complexity of networks?

As areas such as aviation and aerospace, deep-sea exploration, and autonomous driving have undergone rapid advancements, the service scope of wireless networks is constantly expanding. Beyond 5G technology, 6G further integrates satellite communication, AI, and sensing technology (Tang et al., 2024a). Thus, it addresses the limitations of terrain and the surface of the Earth and extends to natural spaces such as space, air, land, and the ocean to achieve "ubiquitous connectivity" worldwide. This makes managing network resources increasingly complex when deploying AI algorithms. Moreover, energy consumption must be considered for the sustainable development of wireless communication.

Based on these issues, we believe that distributed learning for future 6G networks should exhibit five key characteristics, as shown in Fig. 18. The specific features are described as follows:

1. Model following data. With the digitalization of various industries and vigorous development of various B2C terminals, including V2X, abundant data will be generated at the edge of wireless networks in the future. Models are traditionally trained by collecting large amounts of data, which incurs significant

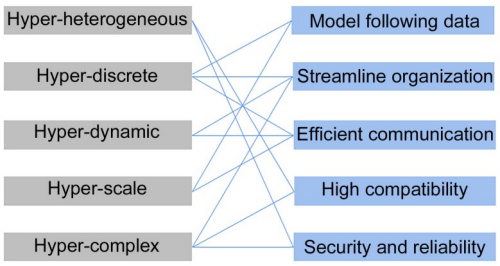


Fig. 18 Characteristics of distributed learning for future 6G networks

transmission overhead and poses privacy risks to users. As carbon reduction, energy conservation, and user privacy security are increasingly emphasized, this approach of having data follow models is no longer suitable for 6G networks. In 6G networks, models should follow data and extract knowledge from data nodes using knowledge distillation.

2. Streamline organization. A set of collaborations participating in distributed learning within a network can be extensive, even across different coverage areas. A traditional network requires only inter-neighbor collaboration. Organizing and coordinating such a large-scale distributed collaboration is a key challenge involved in network management and control capability. Therefore, distributed learning on 6G networks is ideal for lightweight self-organization, which will reduce the management complexity, signaling overhead, and adaptive dynamic adjustment.

3. Efficient communication. Distributed learning deployed on network elements and terminal devices must converge and infer efficiently without frequently exchanging a large number of model gradients and parameters. With increasing model complexity, the number of model parameters has increased. Thus, inefficient distributed learning methods will impact the transmission levels of networks.

4. High compatibility. Devices and terminals with different computing powers from multiple manufacturers constitute a network. Moreover, the observed data dimensions may vary across manufacturers even if the implementation of their model algorithms is different. Thus, distributed learning algorithms running in the 6G network must be compatible with heterogeneous devices' computing power, models, and data categories.

5. Security and reliability. One issue with the model following data learning approach is that it requires transferring and exchanging models among network devices. This raises concerns about the security and reliability of the learning process. As some core model knowledge may not be willing to be shared due to intellectual property, distributed learning in 6G networks must protect model knowledge; this includes allowing each node to share only partial model parameters or distilled knowledge.

These viewpoints indicate a gap in existing distributed learning algorithms, such as federated learning, swarm learning, and split learning, which need to be further optimized within 6G networks. Moreover, new distributed learning paradigms for 6G networks must be explored. The continuous development of AI technology will bring some new challenges for distributed learning. As a breakthrough development of AI techniques, foundation models (FMs) have been fascinatedly applied to many fields represented by chat generative pretrained transformers (ChatGPTs) (Abdullah et al., 2022). In mobile networks, FMs can be used in various fields such as operation, administration, and maintenance (OAM), resource management, and air interface. Hereinafter, to distinguish security and reliability from other industrial applications, we refer to FMs for telecommunication networks as networked generative pretrained transformers (NetGPTs) (Chen et al., 2024).

5.2.5 Heterogeneous hardware for AI computing

5.2.5.1 Heterogeneous hardware deployment and selection for AI computing in AI RAN

Computing hardware plays a vital role in AI applications. Traditional computing relies on general-purpose central processing units (CPUs); however, massive parallel data processing demands of AI exceed CPU capabilities. AI computation often leverages heterogeneous hardware, which is considerably pronounced in RANs due to deployment constraints and low-latency requirements.

In RANs, heterogeneity exists across several levels.

1. Chip level. Heterogeneity involves integrating multiple types of processing units to enhance the computational performance and efficiency of AI models. Using technologies such as Chiplet (https://www.intel.com/content/www/us/en/foundry/chiplets.html), heterogeneous cores and functional modules

can be integrated into a single package such as the DL Boost module in Intel Xeon CPUs. As the advancement of chip manufacturing processes decelerates, this trend toward heterogeneity at the chip level will become more pronounced to meet the increasing demands of AI computing.

2. Node level. Heterogeneous computing involves integrating various types of accelerators within a single server node. Using high-speed interconnect technologies such as peripheral component interconnect express (PCIe) and NVlink, servers can incorporate acceleration cards such as graphics processing units (GPUs), neural processing units (NPUs), or field programmable gate arrays (FPGAs) to enhance the overall computational capabilities. As heterogeneous accelerators can be flexibly combined as required at the node level, this approach is expected to be an important form of heterogeneous hardware in the future.

3. Cross-node level. Nodes within server clusters can be configured using different types of hardware accelerators based on specific application requirements and interconnected via high-speed networks to facilitate AI computing.

Node-level AI heterogeneous hardware for servers must be selected by carefully considering the following factors.

Workload characteristics: applications that are compute-intensive, data-intensive, or require low-latency processing must be considered when selecting hardware.

Flexibility and programmability: task update and adjustment frequency may result in different hardware choices in a server.

Energy efficiency and power consumption: hardware with high energy efficiency must be chosen for RANs.

Interoperability: different types of heterogeneous hardware must be integrated with existing systems with appropriate support for interconnecting protocols.

Along with CPUs, the following common options can be considered when selecting heterogeneous hardware for AI computing.

GPUs: GPUs are renowned for their parallel processing capabilities. With high computational throughput and extensive framework support, GPUs perform admirably in tasks that require massive parallel computations. However, power consumption may pose challenges for certain applications, and GPUs have inherent limitations with latency-sensitive processing.

FPGAs: FPGAs offer unparalleled flexibility and programmability, enabling users to customize hardware accelerators for specific AI tasks. Their parallel processing capabilities make them well-suited for real-time processing applications. However, their programming and optimization may require specialized expertise, and they have lower throughput than GPUs.

Domain-specific architecture (DSA): DSA chips are specialized for neural network computations and provide high performance and energy efficiency. For instance, NPUs that are optimized for deep learning inference provide efficient solutions for inference tasks. However, their lack of flexibility limits the scenarios for DSA solutions such as tensor processing units.

To fully facilitate the AI workflow in RANs and manage network connections during AI training and inference processes, other heterogeneous hardware such as data processing units and network interface cards (NICs) are required.

Specialized servers optimized for AI processing, such as the NVIDIA DGX platform (https://www.nvidia.com/en-us/data-center/dgx-platform/), are commonly employed in large-scale AI cloud clusters, particularly in AI computing centers in the cloud. However, due to highly distributed deployment scenarios and the necessity for flexibility across communication, sensing, and AI capabilities in RANs, traditional CPU-centric servers are adequate for fulfilling the requirements in most cases.

5.2.5.2 Heterogeneous hardware virtualization, management, and scheduling in AI RAN

Cloud technology has become a crucial enabler for seamlessly integrating capabilities such as communication, sensing, and AI deployment in 6G RANs (Tang et al., 2024b). Correspondingly, heterogeneous hardware necessitates virtualization and cloud-native lifecycle management to enable its full utilization in 6G cloud RANs. AI processing demands particularly highlight its necessity to ensure optimal resource allocation and efficient execution of AI algorithms in RANs.

Considering GPUs as an example, virtualization involves two major aspects.

Resource abstraction and encapsulation: the underlying GPU hardware can be abstracted and encapsulated as virtualized instances within the cloud. This virtualization shields the system from the complexities of heterogeneous hardware, enabling unified management and utilization of virtualized resources without the need to delve into hardware details.

Resource sharing and isolation: efficient sharing of GPUs is crucial for 6G cloud RANs, particularly considering the often-limited available resources. Using techniques such as time-sharing and space-sharing, multiple applications in RANs can access GPU resources concurrently. Additionally, isolation is essential in RANs to ensure predictable latency performance. Resource sharing schemes vary in terms of partition types, partition numbers, isolation levels, reconfiguration capabilities, and more. Examples include multi-process service (MPS), time-slicing, and multi-instance GPU (MIG) for NVIDIA GPUs. An appropriate scheme should be selected based on the specific requirements and constraints of each scenario in 6G RANs.

Furthermore, due to the fragmented scenarios and distributed nature of 6G cloud RANs, flexible combination of heterogeneous hardware rather than clustering is commonly observed. Therefore, comprehensive lifecycle management (LCM) of virtualized heterogeneous hardware in a cloud-native manner must be implemented to cater to the various demands of AI applications.

This cloud-based LCM scheme may encompass the following processes.

Resource discovery: it involves identifying the characteristics and capabilities of each virtualized hardware resource within the cloud infrastructure.

Configuration: virtualized hardware resources must be configured based on requirements. For GPUs, configuration parameters may include driver types, partitioning methods, memory allocation methods, and specialized features.

Invocation: it involves requesting access to the required resources and initiating necessary computations. In a cloud environment, Kubernetes (K8s) facilitates this process using YAML files. Applications can efficiently access and use virtualized hardware resources for tasks such as AI inference, AI training, or data analytics.

Release: virtualized hardware resources should be released after tasks are completed. Appropriate release mechanisms can ensure efficient resource utilization and prevent resource wastage.

Cloud-based LCM for heterogeneous hardware can be implemented by designing a heterogeneous accelerator manager (HAM) compatible with K8s. Leveraging K8s and distributed architecture, HAM can effectively manage heterogeneous hardware from various vendors and the corresponding K8s operators or device plugins. This will enable unified management of heterogeneous resources in 6G cloud RANs. Fig. 19 illustrates this concept. However, due to the varying implementations of device plugins by different

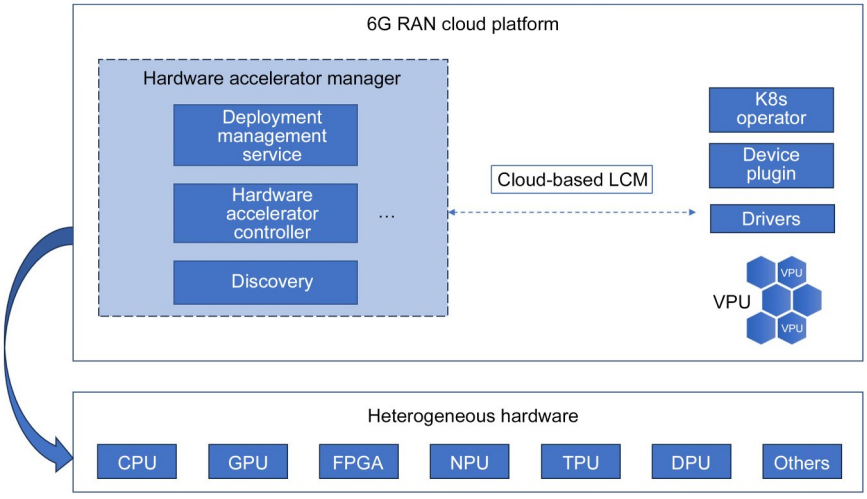


Fig. 19 Cloud-based hardware accelerator manager architecture (RAN: radio access network; LCM: lifecycle management; VPU: virtual processing unit; CPU: central processing unit; GPU: graphics processing unit; FPGA: field programmable gate array; NPU: neural processing unit; TPU: tensor processing unit; DPU: data processing unit)

hardware vendors, designing a unified HAM still requires further studies on alignment with AI computing requirements in RANs.

Resource scheduling is another important issue for heterogeneous hardware in RANs. In traditional cloud environments, heterogeneous resource scheduling primarily relies on K8s' default scheduler. This scheduler prioritizes workload distribution at the node level based on predefined policies such as resource requests and constraints. Although this approach is effective for general resource allocation, it lacks granularity in intranode resource management. In contrast, RANs incorporate new scheduling dimensions such as energy consumption and network topology as well as enable finer-grained resource allocation within nodes. To this end, advanced scheduling mechanisms that can dynamically optimize resource usage at internode and intranode levels are required. This will enhance efficiency and performance in complex, heterogeneous AI computing environments in RANs.

As shown in Fig. 20, a possible scheduling structure can be illustrated as follows:

1. Task intent and task intent interpreter. A task intent defines a high-level goal or an objective of a task, which is then translated into K8s resources. This intent can be provided via configuration or analysis, serving as upstream components that provide essential data for understanding the task's purpose. The task intent interpreter converts these goals into actionable insights, potentially using AI and decision-making modules to optimize task fulfillment. K8s supports various workload types such Deployment, StatefulSet, and DaemonSet, each tailored with specific scheduling and management characteristics.

2. Scheduling intent and scheduling intent interpreter. Scheduling intent specifies constraints and conditions for placing tasks on nodes within the cluster. After a task is represented as a pod or job in K8s, it undergoes scheduling. The scheduling intent interpreter plays a crucial role in processing these constraints, determining the actual requirements for the suitable node placement across the cluster.

3. Cluster scheduler plugin. This plugin processes scheduling intents by considering various constraints to make informed decisions on node placement within the cluster. It can incorporate different scheduling dimensions such as energy consumption and network topology to effectively optimize resource allocation and workload distribution.

4. Node scheduling. In the final stage, the workload is scheduled onto specific nodes within the K8s cluster based on the interpreted scheduling intent. The preferred resource combination within the node is also selected in this stage. This ensures efficient resource utilization and supports dynamic adjustments to meet workload demands across the infrastructure.

In summary, heterogeneous hardware in 6G cloud RANs is essential for AI computing, necessitating

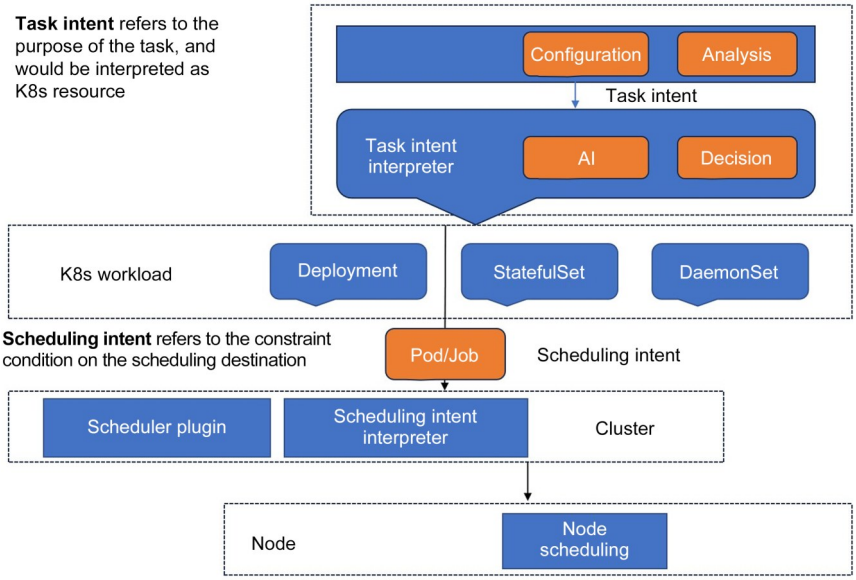


Fig. 20 Cloud-based hardware resource scheduling (AI: artificial intelligence)

careful selection as well as efficient management and scheduling.

5.3 NDTs

5.3.1 Concept of NDTs and design principles

5.3.1.1 NDT concept

An NDT is an emerging paradigm founded by integrating data and models for digitizing physical networks into virtual models (Almasan et al., 2022). By employing real-time data and simulation analysis, NDT achieves synchronization and feedback with physical networks. This enhances capabilities such as virtual-to-physical mapping, intelligent decision-making, simulation and pre-validation, and closed-loop optimization lacking in physical networks. As a foundational framework that supports unified planning, orchestration, and management of 6G networks, NDT facilitates comprehensive upgrades in network planning, construction, maintenance, optimization, and operation. Moreover, it empowers networks to achieve a high level of autonomy characterized by "zero-touch, zero-wait, zero-fault" operation (Zhang LF et al., 2022).

5.3.1.2 Design principles

NDT must obey the following design principles:

1. On-demand and agile. NDT must have intent comprehension capabilities to understand users' vague intents using techniques such as LLMs. NDT must break down these intents into sequences of tasks and flexibly orchestrate and allocate resources to complete them.

2. Autonomous and intelligent. NDT must have autonomous intelligence management. It must use AI-based intelligent algorithms and tools for digital twin model lifecycle management, ensuring closed-loop deviation control between the digital and physical networks and maintaining its performance (Institute CMCCR, 2022).

3. Real-time and native. NDT should serve as an inherent native entity of future 6G networks and be considered in the design phase rather than retrofitted as additional functions. As a native mechanism, NDT can seamlessly integrate with other functions within 6G networks. It can thus considerably reduce overhead and promote real-time interaction between physical and digital networks.

4. Compatible and open. The interfaces and models of NDT should be fully compatible and invoked by other noncommunication systems, and should have the capacity to be integrated into a larger digital twin world model and support application upgrades and the emergence of new applications.

5. Cloud–edge collaborative. NDT should fully use the cloud's abundant computing power and edge's low-latency advantages to wisely allocate responsibilities between the cloud and edge and collaboratively complete tasks via distributed cooperation among multiple nodes.

6. Security and privacy. NDT should protect user privacy data and resist malicious attacks. Ensuring the fidelity of NDT to its physical counterparts demands rigorous security for safeguarding it against data manipulation and integrity breaches. With its significant role in managing personal data, privacy measures are indispensable to prevent leaks, particularly during data collection and processing. The network architecture must contemplate mechanisms for data security and privacy protection from the design phase.

5.3.2 NDT framework for 6G network autonomy

NDT achieves real-time interaction and virtual–real mapping with a physical network by accurately modeling the physical network in multiple dimensions, including time and space, thereby enhancing system-level simulation, optimization, pre-verification, and control capabilities that are lacking in the physical network. High-level autonomy is an important feature of 6G mobile networks (Zhang LF et al., 2022), and NDT is considered the most potential technology to realize 6G autonomous networks. To meet the requirements of high-level intelligent and flexible autonomy for 6G networks, a layered cross-domain collaboration framework of "centralized control + distributed autonomy" is innovatively proposed (Fig. 21). The global-domain NDT entity supports end-to-end autonomous requirements, whereas local domains such as the access domain, transmission domain, and core domain support sub-domain autonomous requirements.

China Mobile has proposed a novel architecture termed as "3 entities, 4 layers, 5 planes (3–4–5)," where "5 planes" include the control, user, data, computing, and security planes and "3 entities" include the

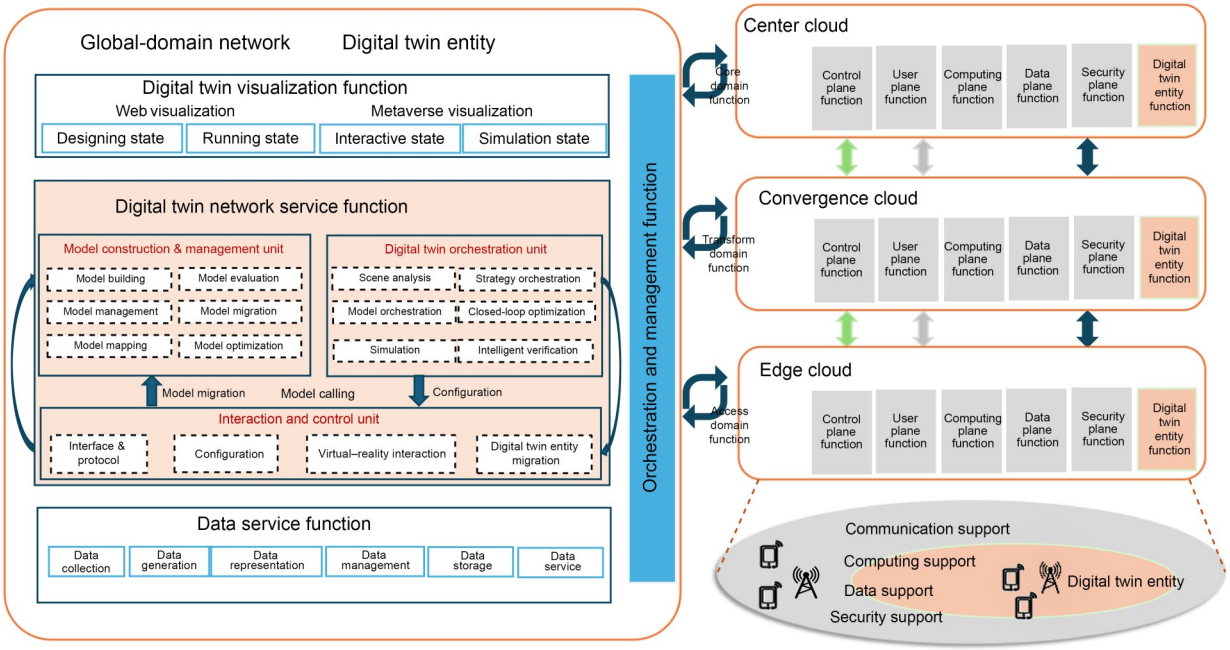


Fig. 21 Network digital twin (NDT) framework

network, digital twin, and management orchestration entities (Bhat and Alqahtani, 2021; Raj et al., 2023). The proposed NDT framework is further designed based on the above architecture. Each local domain has "5 planes" and a sub-domain digital twin entity. These planes are invoked to support the realization of digital twin functionalities. Moreover, the traditional control plane and user plane ensure communication between digital twin entities. The data plane provides data to the twin entities to update the twin models and ensure real-time interaction with the physical network. The computing plane provides computing resources to guarantee the stable operation of digital twin entities and the calculation of twin tasks. The security plane protects the digital twin from external attacks and ensures the normal operation of digital twin entities.

The end-to-end digital twin entity has four functions: data service function, NDT service function, orchestration and management function, and digital twin visualization function. The data service function builds unified representation data based on the data collected from the "data plane" to provide services for digital twin modeling. It also comprises other functions such as data augmentation, data storage, and data management. Orchestration and management function is used mainly to parse intents and perform business logical orchestration based on intents and schedule domain twin entities to complete twinning tasks, thus ensuring closed-loop iteration. The NDT service function is the core of the digital twin entity, which is used for twin modeling to achieve simulation and pre-verification. Three functional units, namely the model construction and management unit, digital twin orchestration unit, and interaction and control unit, have been therefore innovatively proposed to build the core functions of the digital twin entity. The accuracy and granularity of these units determine the twinning and mapping capabilities of the DTN to the physical network.

1. Model construction and management unit. This unit has main functions such as model construction, iteration, and management. The construction of an NDT model is a comprehensive representation of the physical process that comprises the entire lifecycle of network planning, construction, maintenance, optimization, and operation on multiple temporal and spatial scales. This unit has multidimensional, precise, and fine-grained modeling capabilities for the physical, behavior, rule, and decision models of a physical network. The constructed digital twin entity can interact with the physical network in real time to perform activities such as data perception, accurate modeling, real-time dynamic simulation, multidimensional pre-verification,

virtual and real mapping, closed-loop optimization, and intelligent decision-making.

2. Digital twin orchestration unit. This unit has functions such as on-demand orchestration, simulation and pre-verification, and policy generation. Tailored to various scenarios, this unit orchestrates multiple models to form a digital twin environment based on the network topology relationship. Based on this digital twin environment, data can be generated for AI training to enhance AI model generalization. It can also serve as an interactive environment for AI training, thereby improving the convergence speed of AI models. It can support the pre-validation of AI models, thereby enhancing the feasibility of network policy and reducing the risk of failures.

3. Interaction and control unit. This unit has functions such as digital twin protocol and interface, configuration, virtual and real interaction, and digital twin migration. The policy verified by the digital twin entity can be configured to the physical network via a virtual and real interactive control interface. Digital twin entity migration is also supported by this unit, allowing the transfer of functionalities between the end-to-end digital twin entity and domain digital twin entity, facilitating resource reuse.

5.3.3 Key technologies for NDTs

5.3.3.1 Data collection and generation

NDT hinges on data that must meet the performance requirements for synchronizing the physical and digital domains. It has the following supporting technologies for data acquisition and representation:

1. Data collection technology

NDT data are large and rich and have high granularity. To ensure the synchronization between DTN and the actual network, the NDT should accurately perceive massive heterogeneous data based on the specific requirements of network autonomy scenarios. It should dynamically adjust the scale of collected data according to changes in the network to avoid data redundancy and large data overhead. 6G network data have the characteristics of multi-source heterogeneity and loose organization structure, making it challenging to efficiently obtain data from various application scenarios (Raj et al., 2023).

Data lightweighting technology can reduce the volume of collected data. Knowledge graphs are used to analyze the correlation between extracted features based on the endogenous relationship of the network protocol as well as the performance of the current network, identify the required data types, and realize data collection on demand. Real-time performance of data can be supported by setting a data update mechanism that integrates multiple update modes, such as scheduled and triggered updates, and establishing a real-time data collection protocol. A data accuracy guarantee mechanism (such as cloud–edge data collaborative calibration (Zhu et al., 2021) and resampling) must be established and appropriate data processing technologies (such as data filling, data prediction, and outlier processing) must be adopted to obtain high-precision data. Although a standardized data format is preferred, the vast number of use cases makes it impractical to define a specific data format for each scenario. Therefore, flexibility must be ensured for nonstandardized data formats. Thus, standardized data formats and data interfaces are required for data from different manufacturers to establish a unified data collection mechanism. A data dictionary can be used to manage the data to be collected.

The following design principles can be used for the volume and time scale of data collection and transmission:

(1) Meet the synchronization and tracking requirements of digital twins in various 6G autonomous network states and distinct domains. These 6G autonomous network states include the operation, O&M, and business states, whereas the distinct domains include the RAN, core, and transmission domains. Each possessing demands varying granularities and time scales. DTN should adjust the granularity and time scale of data collection based on different states and domains, for instance, tracking traffic tides over hours, RRC events within milliseconds, and L1 functions within nanoseconds.

(2) Meet the real-time requirements of functions. The acquisition time granularity must not be greater than the time interval of the function implementation, such as the pre-verification function.

(3) Meet the basic data requirements of the functions. Different types of data must be collected for different scenarios and volumes for different functions. For example, the algorithm training and update to realize the weight optimization decision of MIMO

scenarios should track the coverage quality and user behavior data stored for days or even weeks.

(4) To meet the needs, the burden caused by excessive data collection and transmission to the network should be avoided.

2. Data generation technology

DTN obtains large amounts of data via data collection, thereby burdening data transmission. Data privacy needs of users and vendors pose obstacles to data collection. These abundant data can be generated using data generation technology to reduce the collection overhead. Datasets on faults that are usually difficult to obtain on existing networks can also be generated and expanded through NDT to improve the accuracy and generalization of the fault identification and location algorithm. Applicable data generation technologies include data generation based on ML (such as data augmentation based on generative adversarial network (He WL et al., 2023; Hui et al., 2023) or generative pre-trained transformer (GPT)) and data generation based on digital twin scenarios.

3. Data storage technology

DTNs collect data via cross-vendor, cross-system, and cross-domain, which requires more flexible data storage technology for support. Distributed data deployment can store data nearby and reduce data transmission delays. Historical data are stored and cleaned regularly to increase data reusability. AI-based data mining (such as principal component analysis (PCA)) can reduce the storage space by transforming high-dimensional data into low-dimensional subspaces (TG3, 2023).

5.3.3.2 Digital twin model construction

NDT modeling for 6G RAN comprises two layers. The first layer involves modeling various network components such as network elements, user devices, wireless channels, and services to facilitate inference and automatic management of network parts. The second layer integrates these different component models into a comprehensive network model to achieve full lifecycle automation for the entire RAN system (Fig. 22).

1. Modeling for network components

Several critical network components must be modeled in NDTs for RANs to facilitate network monitoring and inference. These include network elements, mobile devices, wireless channels, services, and more.

(1) Modeling for network elements. This involves digitally replicating physical network elements and their relationships in a comprehensive, low-cost, and intelligent manner. For modeling a physical network element for NDTs such as BSs, relays, UPFs, and edge computing nodes, its properties and functions

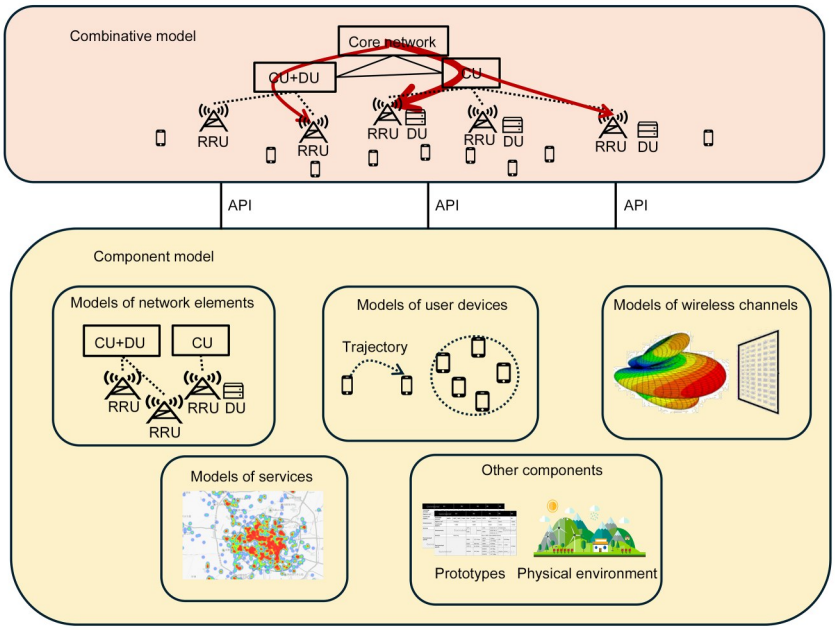


Fig. 22 Model construction for NDTs (NDT: network digital twin; CU: centralized unit; DU: distributed unit; RRU: remote radio unit; API: application programming interface)

must be presented with sufficient intelligence and accuracy for state inference and predictive fault detection while minimizing computational complexity. It must represent complex relationships between homogeneous and heterogeneous elements, including connection topology, master–slave relationships, collaborative relationships, and redundancy. This relationship modeling should be comprehensive, scalable, automatic, and cost-effective. Knowledge graphs, graph neural networks (GNNs), and generative AI are typically used for modeling network elements. Knowledge graphs can efficiently represent and store attributes and relationships between different elements, thereby improving the efficiency of data acquisition and utilization (Zhang SY et al., 2023). In more complex network environments, neural networks such as long short-term memory and GNNs can effectively extract temporal and spatial features between different types of network nodes and capture the relationships between nodes (Ferriol-Galmés et al., 2022). Finally, generative AI can automatically learn complex relationships between various elements and play a critical role in network modeling in case of low training costs.

(2) Modeling for mobile devices. This involves digitally replicating the properties and behaviors of specific users/terminals or groups of users for NDTs. The model needs to efficiently store, analyze, and infer various properties and behaviors for each user. These properties include device versions, battery strategies, user preferences, credibility, and service requirements, whereas behaviors encompass user trajectory, browsing habits, and other factors. This facilitates user-specific QoS/quality of experience (QoE) optimization in NDT automation. However, user privacy concerns make acquiring data for modeling challenging. To address this issue, transfer learning can be used to first train models on large-scale universal datasets and then fine-tune them to adapt to the limited data of individual users. For modeling a group of users, the system needs to infer group requirements (e.g., density scenarios in large stadiums) and behaviors (e.g., tide phenomena). These parameters are then used to optimize network configurations, thereby enhancing network usage efficiency and service quality in NDTs. To overcome the associated challenges including data heterogeneity, balancing personalization with generalization, and maintaining data privacy, data standardization and ML algorithms can be used for handling and interpreting diverse data. Moreover, federated learning can be leveraged to train models using data from multiple devices without sharing user data, thus protecting privacy while ensuring generalization. Generative AI can be used to generate user behavior data in various scenarios, thereby enriching the dataset while adhering to stringent privacy policies.

(3) Modeling for wireless channels. This involves modeling channel characteristics across various time–frequency–space resources in actual physical environments within 6G networks. This process can predict channel fading and proactively adapt to communication requirements using reconfigurable intelligent surfaces, thus achieving automatic channel optimization in 6G NDTs. However, the modeling process is challenging due to the multidimensional coupling of time–frequency–space resources, stringent accuracy requirements, denser deployment of access points, and extremely large-scale MIMO environments in 6G networks. To address these challenges, computer vision can be leveraged to automatically reconstruct 3D propagation environments for specific physical environments. Moreover, cluster-nuclei ideology based on 3D geometrical and electric field calculations can be used to efficiently and accurately model 6G channels for various scenarios (Yu et al., 2022). Propagation environment semantics can also be used to facilitate task-oriented channel modeling and predictions (Sun et al., 2023).

(4) Modeling for services. This refers to the digital replication of various edge services to express their properties and behaviors, e.g., traffic distributions and resource requirements. This is a fundamental component in DTN for network planning and optimization based on accurate traffic predictions and simulations across different areas or nodes over short or long time spans. However, as new services (such as virtual reality, holographic communication, and vehicle networks) have emerged in 6G networks, developing models for different services automatically, on demand, and at low cost becomes challenging. This issue can be addressed using ML that effectively learns different patterns of various services based on extensive data. Generative AI can also provide more data for model training, thereby enhancing the accuracy and reliability of these models.

2. Combinative modeling

The second layer integrates the models of various network components to create a more comprehensive network model, thus enabling high-level inference for different stages of network management in NDTs. For instance, during the network optimization stage, this integrated model can analyze fine-grained relationships (such as topology) and states (such as resources and fault probabilities) of different physical network elements, predicted wireless channels, future user behaviors, and estimated traffic loads. Based on this analysis, it can predict network performance and fine-tune network configurations from multiple perspectives, such as network-wide or link-specific views.

To this end, a platform must be established for organizing different component models via APIs. This platform can conduct system-wide inferences by intelligently combining the inferences from various component models in DTN. Specifically, the platform can be built on large base models by leveraging their ability to merge knowledge from different fields. Such a platform can also prompt AI agents to collaborate effectively to achieve network performance goals.

The agent can gain knowledge by applying supervised learning and reinforcement learning (RL). In supervised learning, massive historical network data contain rich expert settings and module iterations under various network conditions. It can be used to build large datasets for fine-tuning large FMs in NDT agents and training API calls between different agents. In RL, NDT agents can interact with the pre-validation environment in DTN to obtain estimated network performance, which can be used for evaluating agent behaviors during the training process in RL. Specifically, as the pre-validation environment can train NDT agents ahead of time, the concern for timeliness is negligible. As FMs have hallucinations, even if RL (based agent) converges in the pre-validation environment of NDTs, they must compare the agent's performance with that of traditional expert strategies (via channel state information (CSI)) before they are adopted in the physical world.

5.3.3.3 Digital twin orchestration and management

DT orchestration and management are key technologies for optimizing network strategies, improving the deployment of AI applications, and enhancing the network O&M efficiency (Lu et al., 2021; Ferriol-Galmés et al., 2022; Raj et al., 2023). The orchestration and management design of DTs should support an operator to flexibly instantiate and use the DT system functionalities.

1. NDT orchestration

NDT orchestration refers to the intelligent orchestration of DT models to achieve internal closed-loop optimization control; it makes NDT flexible, scalable, and self-evolving, and enables NDT to adjust execution policies based on physical demands as well as achieve superior pre-validation performance (Deng et al., 2022; Fig. 23).

The closed-loop orchestration of NDT involves orchestrating scene-oriented virtual environments, simulation models, and intelligent strategies using the constructed DT models to achieve closed-loop simulation verification and iterative optimization of strategies. A complete internal closed-loop process is fulfilled by sequentially executing scene orchestration, model orchestration, and policy orchestration.

Scene orchestration uses intent parsing for scene recognition and exploring and representing the network elements and topological relationships involved in the scene. The required network elements are retrieved from the model library for deployment, and digital twin scenes are orchestrated using a topological representation of relationships.

Model orchestration selects simulation models corresponding to the network element nodes based on the virtual network environment and constructs a network simulation engine. The accuracy and consistency of the NDT are ensured by combining the involved models, jointly optimizing multidimensional models, aligning parameters with the physical network, and intelligently verifying the overall twin environment.

Based on the constructed and orchestrated NDT, the intelligent strategies generated by AI are orchestrated for policy pre-validation. If the policy feasibility is not satisfactory, iterative optimization must be continued until the simulation results are met and the configuration is issued.

2. DTN management

DTN management ensures the lifecycle management of the DTN, including comprehensively characterizing the physical processes of network organization, construction, maintenance, optimization, and

operation from multiple temporal and spatial scales such as model construction, iteration, management, and driving (Zhu et al., 2021). By managing the real-time mapping and interaction between DTNs and physical networks, processes such as global data perception, efficient modeling, real-time simulation, multidimensional verification, closed-loop optimization, intelligent decision-making, and holographic visualization can be performed. A typical lifecycle management is shown in Fig. 24.

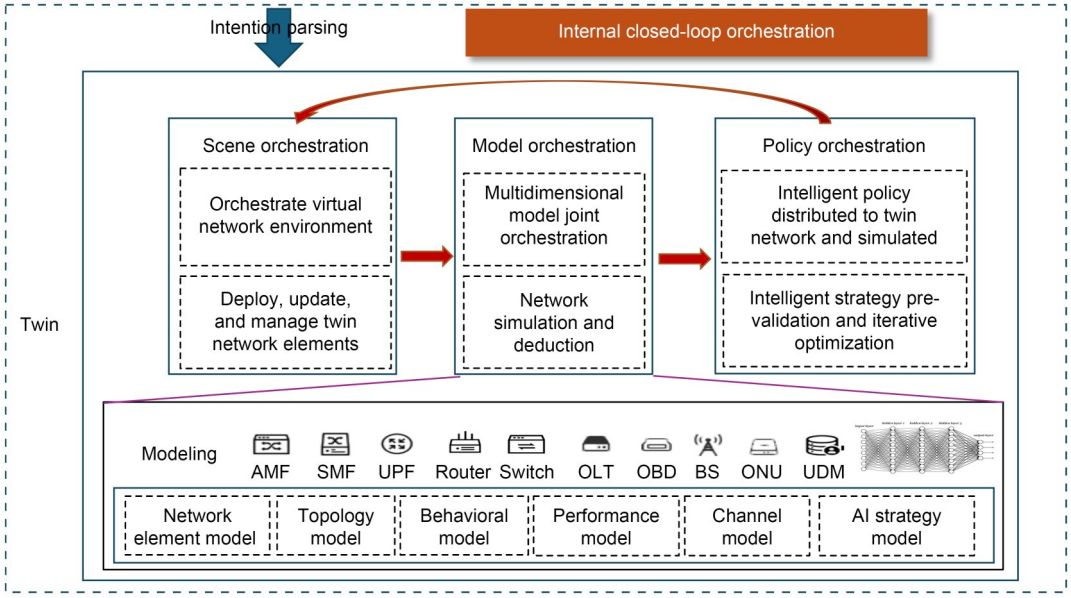


Fig. 23 Closed-loop orchestration of NDTs (NDT: network digital twin; AMF: access and mobility management function; SMF: session management function; UPF: user plane function; OLT: optical line terminal; OBD: on-board diagnostics; BS: base station; ONU: optical network unit; UDM: unified data management; AI: artificial intelligence)

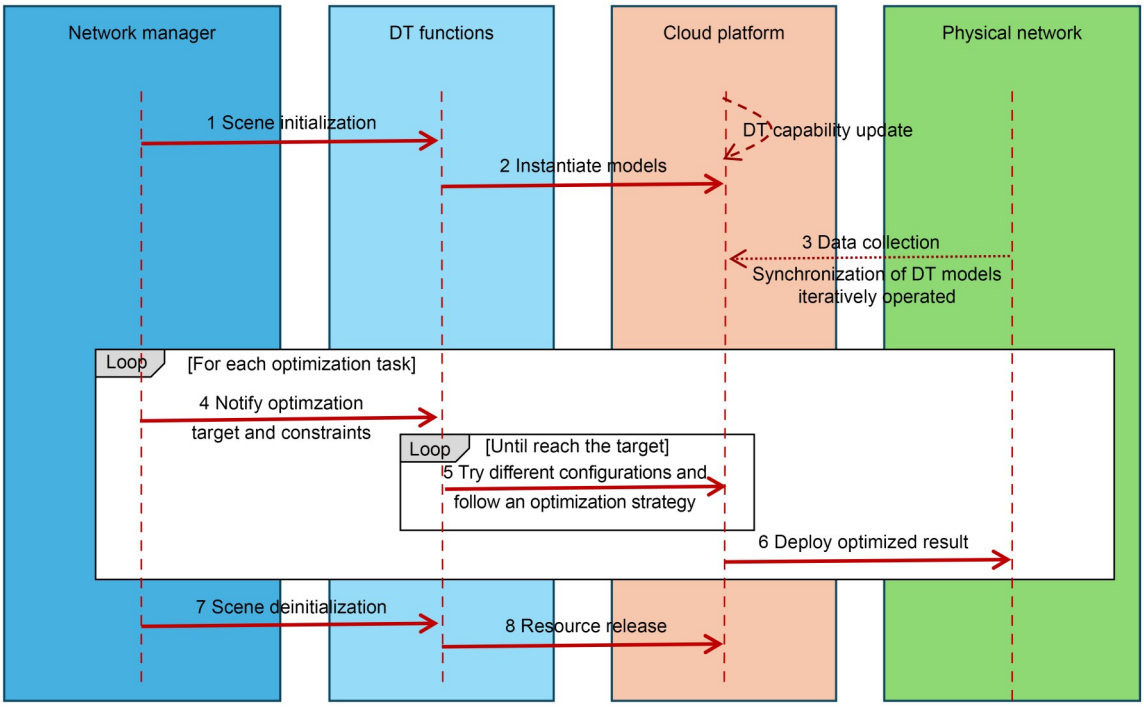


Fig. 24 Example of DTN lifecycle management (DTN: digital twin network; DT: digital twin)

3. Registration and updating of digital twin capability

The computing power required for the DTN may be distributed in different network nodes. A DTN element may colocate with the actual network element or be instantiated in a cloud platform. If a network node is prepared to support a digital twin service, it should report its capability to the cloud platform or the digital twin service controller for registration. In case of changes in the capability, the network node should report its new ability to assist the orchestration in the controller.

4. Initialization and destruction of digital twin scene

The digital twin service controller must instruct the cloud platform or the corresponding network nodes to initialize digital twin scenes. One scene may contain digital twins for a CN, BSs, user terminals, and wireless channels. The generation of network elements may rely on the virtualization and container technology such as virtual box, composer, and Docker. If all the digital twin tasks corresponding to the scene are performed, the controller should be able to instruct network element destruction.

5. Synchronization of digital twin scene

Due to time variations in a wireless network, the digital twin part must be continuously fine-tuned to adapt to changes in the physical network. The time variation may be due to the changes of the wireless channel and the movement of UE. Thus, the digital twin scene needs to continuously gather corresponding data from the real network, which can be transmitted via the data plane.

6. Performing digital twin tasks

DTN can evaluate new network parameters and provide advice for network optimization. The service controller can choose the optimized parameters and a target. DTN should automatically optimize the parameters via iterative evaluation using Bayesian optimization or RL methods. The optimization results will be deployed in the physical network.

5.3.3.4 Intent-driven NDTs

Intent-driven management involves guiding and automating network operations by defining high-level business goals or desired outcomes. Unlike traditional network management, intent-driven management emphasizes “what to do” rather than “how to do it.” Administrators need only to define the desired effect to be achieved by the network, and the intelligent system automatically handles the specific implementation. As the enabling technology of NDTs, intent-driven management plays a crucial role in autonomous intent parsing, policy generation, closed-loop operation, iterative optimization, and policy validation (Fig. 25).

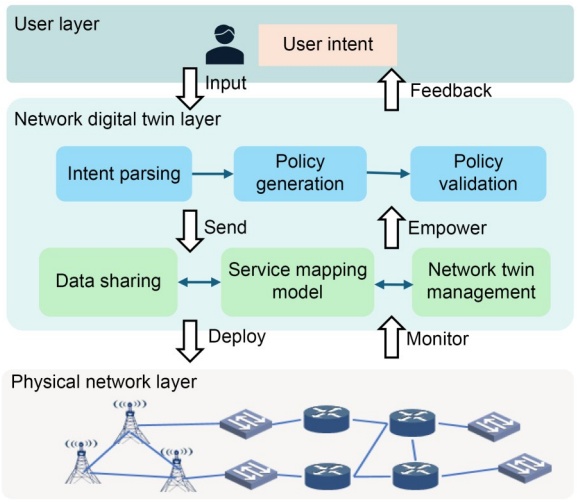


Fig. 25 Primary workflow of intent-driven network digital twins (NDTs)

1. Autonomous intent parsing. An intent-driven NDT considerably enhances network management autonomy by automating intent parsing. This automation ensures that high-level business goals are accurately translated into network configurations without manual intervention, thereby reducing the risk of errors and improving efficiency (Mehmood et al., 2023). Intent parsing involves interpreting and understanding high-level intents expressed in a natural language or abstract terms. Natural language processing techniques are used in intent parsing to analyze and understand the natural language inputs from administrators (Patwardhan et al., 2023). This allows the system to extract relevant information and identify specific goals or desired outcomes. Using intent recognition, the intent type can be identified and then mapped to predefined network actions or configurations. This involves using trained AI models to identify various intents based on context and content. Semantic analysis is performed to understand the context and

relationships between different parts of the intent to ensure accuracy. This helps resolve ambiguities and ensure that the intent is fully understood.

2. Policy generation. After an intent is parsed, policies that can be implemented within the network to achieve the desired outcomes are generated. The system first maps high-level intents to specific network actions or configurations by identifying the necessary changes in network parameters, routing protocols, security settings, and other functions. Policies are then defined in a structured format that can be easily implemented by network devices. These policies include rules and conditions that govern network behavior, such as access control lists, QoS settings, and traffic management rules. Finally, scripts are generated to configure network devices according to the defined policies for automating the implementation process. These scripts ensure that the changes are applied consistently across the network.

3. Closed-loop operation. The intent-driven approach guarantees that the DTN operates within a closed-loop system. It ensures that the network remains aligned with user intents by continuously validating and adjusting policies based on real-time feedback from the physical network. This closed-loop mechanism enhances reliability and responsiveness and enables the network to adapt dynamically to changing conditions (Khan et al., 2022).

4. Iterative optimization. A critical advantage of an intent-driven NDT is its capability for iterative optimization. By leveraging continuous feedback, the system can iteratively refine its operations and improve its performance over time. This iterative process ensures that the NDT meets and exceeds initial performance expectations, adapting to new challenges and opportunities.

5. Policy validation. Policy validation is crucial to ensure that the generated policies achieve the intended outcomes without causing disruptions or conflicts in the network. To this end, policies are simulated and tested in a digital twin environment before deploying them in a live network. This helps identify potential issues and ensures that the policies work as expected. Conflict detection is designed to check conflicts between new policies and existing configurations. Conflict detection algorithms are used to identify and resolve any overlapping or contradictory rules. Policies are verified against regulatory and organizational compliance requirements. This ensures that the network remains secure and compliant with relevant standards and regulations.

5.3.4 Use case of NDTs

In the context of future 6G wireless communication systems, an NDT will play a pivotal role across various domains such as AI training, AI pre-verification, parameter pre-verification, and performance verification. Herein, the role of NDTs is illustrated via an example of the L2 downlink scheduler in BSs (Fig. 26). This scheduler can support multi-user communication scheduling in time, frequency, and space domains, along with individualized link adaptation for each user. Due to the complexity of multi-user spatial multiplexing, AI models are employed to handle the spatial pairing of different users and select the optimal modulation and coding scheme (MCS) for various pairings while the remaining modules and components continue to use conventional algorithms.

1. AI training. A network-level DTN is initially constructed in the cloud, which comprises digital twin modeling of typical environments, user distributions, and other BS algorithms and components. This DTN is further structured as a reinforcement learning environment with well-defined reward functions (Tao et al., 2023) such as functions based on aggregated throughput of paired users, thereby enabling AI models to learn optimal multi-user multi-input multi-output (MU-MIMO) user pairing and MCS selection criteria. Due to variations in user distributions and channel characteristics across different cells, a smaller-scale DTN tailored to specific field conditions is deployed at the network edge. This facilitates fine-tuning of the initial model to obtain AI models adapted to particular field conditions.

2. AI pre-verification. As inherent uncertainties are associated with AI decisions due to factors such as data, models, and environmental variations, AI models must be pre-verified. After obtaining AI models for MU-MIMO user pairing and MCS selection, these models are deployed within edge DTNs to assess their performance. Edge DTNs replicate specific field conditions, including real BS hardware and software versions, enabling the evaluation of AI model results on real-world performance and issues. This pre-verification

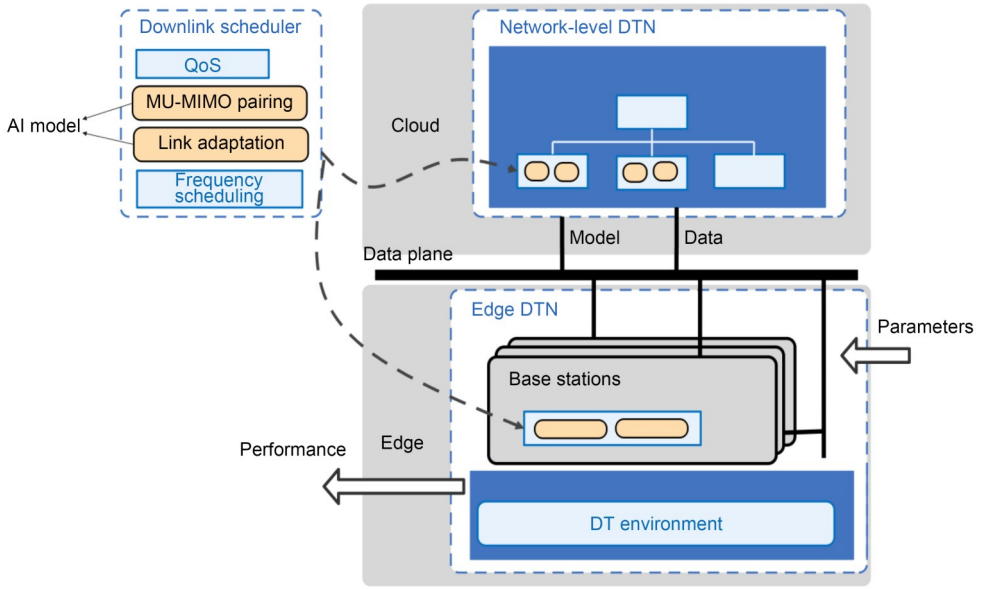


Fig. 26 Use case of NDTs (NDT: network digital twin; QoS: quality of service; MU-MIMO: multi-user multi-input multi-output; AI: artificial intelligence; DTN: digital twin network; DT: digital twin)

helps detect anomalous MU-MIMO user pairings and MCS selections, isolating their impact on the physical network.

3. Parameter pre-verification. Modern communication systems involve many interrelated parameters, which makes analytical modeling challenging. For instance, parameters influencing fairness factors and RAN slicing in L2 schedulers affect L2 performance correlation with other L2 parameters. They also impact the overall system behavior and performance, including L1 and L3 settings. Analysis alone cannot be used to effectively pre-assess the effects of parameter modifications on the system behavior and performance. Edge DTNs facilitate configuring parameter changes and combinations in advance, observing system outputs, and evaluating comparative results while mitigating the influence of random factors. By coupling DTNs with AI-based parameter optimization algorithms (Eriksson et al., 2019), vast parameter space can be efficiently explored to identify optimal parameter combinations for different field conditions and considerably reduce the time required for on-site test and evaluation iteration.

4. Performance verification. System performance is a culmination of various factors, including design, parameter settings, algorithm performance, and hardware/software versions, despite optimized and well-designed L2 scheduler algorithms. Mathematical calculations and analyses are insufficient for assessing performance because of several influencing factors. Edge DTNs, with detailed modeling of field environments and user behavior/distribution, provide a virtual environment for precise performance evaluation of real products' software and hardware. This evaluation comprises metrics such as MU-MIMO user pairing ratios and performance gains in specific field conditions. Furthermore, a hybrid virtual–physical future scenario can be constructed by configuring and modifying digital twin models manually. For instance, by modeling idle-time harbor scenarios and combining them with analysis and forecasting of future harbor operations, future busy harbor scenarios can be modeled. This will enable foreseeing the performance of MU-MIMO scheduling. This approach enables early assessment of future scenarios, facilitating advanced planning of product capability requirements.

6 Prototype and trial

6.1 DTN prototype design and key technologies

Based on the network element models and topological models, a unified database representation is used for the on-demand construction of a foundational network model. This facilitates virtual–real mapping between the DTN and physical network. Digital twin

management uses digital threads to feed information from each step of network operations into the model for optimization, prediction, and guidance. Closed-loop feedback analyzes implementation results and iteratively refines the model to maintain deviations within acceptable tolerances. This holistic digital twin management spans the entire network lifecycle—from simulation to decommissioning—ensuring continuous optimization and management of end-to-end network segments (Wang SF et al., 2023a).

The DTN prototype framework, designed for highly complex enterprise-level applications, adopts a layered and modular approach to ensure scalability, flexibility, and maintainability (Fig. 27). The front-end layer leverages HTML5 technologies, dynamic jQuery scripts, and the Vue framework to create a rich user interface and uses Three.js and Cesium for sophisticated 3D rendering and data visualization. The back-end layer employs a microservice architecture, leveraging SpringBoot and Docker containerization for efficient service deployment and management. The middleware layer coordinates data flow, processing, and communication between different services. The core service layer focuses on AI and computational capabilities, emphasizing the data processing and intelligent analysis capabilities of the system. The data storage layer employs a combination of SQL and NoSQL database technologies to provide flexible data management and optimized storage solutions. This prototype, incorporating virtualization and cloud service technologies, demonstrates deep consideration for resource optimization and future technological expansion. It aims to support modern enterprises in addressing the challenges of big data and intelligent transformation. An example implementation of a DTN prototype is shown in Fig. 28.

1. High-throughput data acquisition

Comprehensive data collection is challenged by the vast scale of the physical network, diverse device forms, and dynamic traffic information (Wang SF et al., 2023b). Therefore, an on-demand acquisition model must be adopted which is tailored to meet specific application requirements. Data types, frequencies, and methods should align with the applications of the NDT, ensuring both comprehensiveness and efficiency.

When building data models for specific network applications, the required data can be efficiently retrieved from the data-sharing repository of the DTN. This repository serves as the single source of truth, storing vast amounts of historical and real-time network data. By integrating these data into a unified environment, the DTN provides a standardized interface and services for data modeling. Data acquisition software supporting MySQL, MongoDB, Redis, and AntDB databases will be developed. Due to the large volume, variety, and high speed of network data, a combination of diverse storage and service technologies will be employed to construct the data-sharing repository.

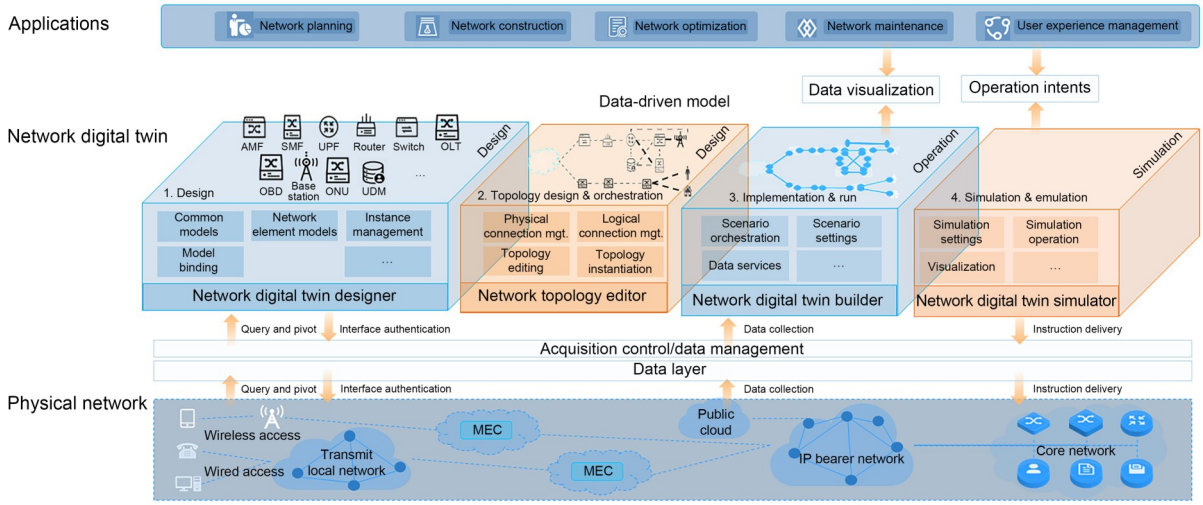


Fig. 27 Framework design of a DTN prototype (DTN: digital twin network; AMF: access and mobility management function; SMF: session management function; UPF: user plane function; OLT: optical line terminal; OBD: on-board diagnostics; ONU: optical network unit; UDM: unified data management; mgt.: management; MEC: multi-access edge computing; IP: Internet protocol)

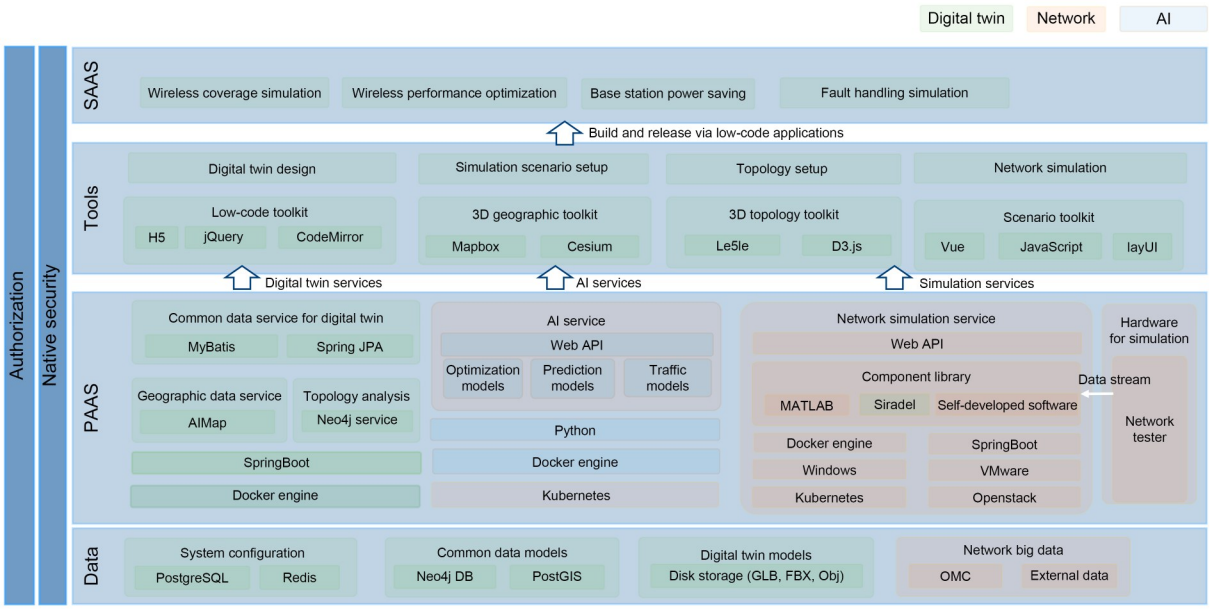


Fig. 28 Implementation of the DTN prototype (DTN: digital twin network; AI: artificial intelligence; SAAS: Software as a Service; API: application programming interface; OMC: operation and maintenance center)

2. High-precision digital twin modeling

DTN modeling for a wireless network primarily comprises individual device modeling and application scenario functional modeling.

Individual device modeling provides the foundation for the consistent representation of large-scale network data. In particular, ontology-based approaches can be leveraged to represent entities. Using physical device information, environmental information, topological node information, network link information, container/virtual machine information, and network element configuration information, individual digital twin models can be developed for communication network devices and logical network elements.

Functional modeling, tailored to the requirements of actual NFs, uses diverse functional modules throughout the lifecycle to achieve dynamic and evolving network reasoning and decision-making. These functional models can be constructed and expanded across multiple dimensions based on the needs of various network applications.

3. High-fidelity digital twin rendering

DTN visualization requires support for different levels of detail in network element models and scenario models. Based on the specific application, an appropriate level of visualization fidelity should be chosen to construct a corresponding virtual environment. Visualization capabilities must encompass an entire network lifecycle from micro-scenarios to large-scale city-level scenarios. This technology enables a high-fidelity representation of the actual network driven by digital threads. Network visualization can be modified using these digital threads, which results in dynamic visualization capabilities. The goal is to create a parallel digital twin world of the physical network, from macro to micro, indoor to outdoor, personalized to large-scale, and network-wide to individual devices.

Network visualization technologies offer two key benefits: they assist users in understanding the internal structure of the network and facilitate the discovery of valuable information hidden within the network. However, DTN visualization presents challenges due to the network scale and the requirement for real-time virtual–real mapping.

6.2 Prototype performance validation for energy-saving scenarios

Cloud-RAN (C-RAN) comprises network elements such as CU, DU, and active antenna unit operating on general-purpose x86 server platforms. A BBU power saving workflow is shown in Fig. 29. Data collected by the BBU, including physical resource block (PRB) utilization, amount of UE, reference signal received power (RSRP), data traffic, CPU usage,

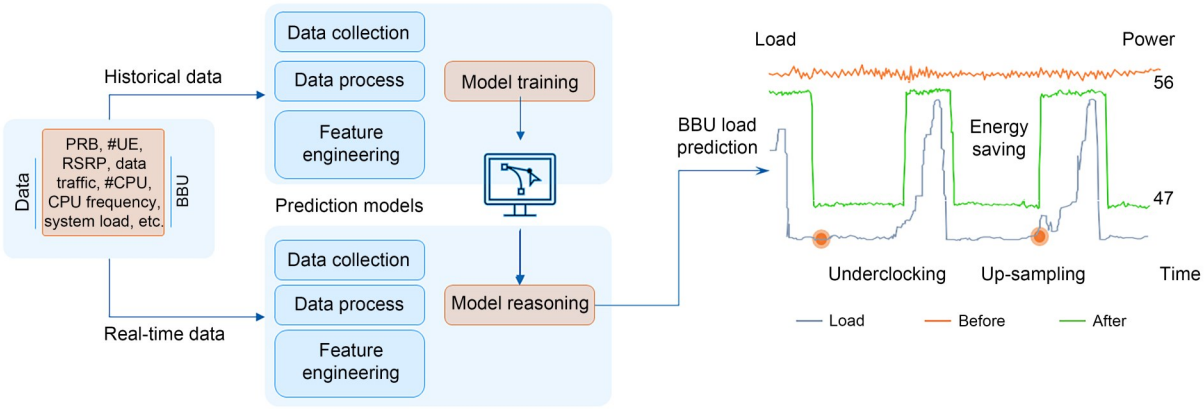


Fig. 29 Working flow for power saving in BBU (BBU: baseband unit; PRB: physical resource block; UE: user equipment; RSRP: reference signal received power; CPU: central processing unit). References to color refer to the online version of this figure

CPU frequency, and system load, are used as the input for AI model training. Based on the predicted traffic patterns, the AI model dynamically adjusts BBU computing resource allocation to achieve energy savings by adapting to the BBU workload.

1. Test configuration

To evaluate the impact of traffic dynamics on energy efficiency, tests were conducted under various traffic scenarios; these included extended periods of inactivity, short periods of inactivity, low traffic load, and high traffic load. The following cloud computing resources (Table 3) were used for this test.

Table 3 Test configuration

Item	Description
CPU	Intel® Xeon® D-2177NT CPU @1.90 GHz
Number of CPU cores	28
On-line CPU list	0–27
Number of threads per core	2
Number of cores per socket	14
Number of sockets	1

2. Performance results

Under high traffic load scenarios, where BBU computing resources are in high demand, enabling the energy-saving mechanism did not considerably reduce the energy consumption. However, during intermittent traffic or under low traffic load conditions, the proposed approach achieved an overall energy reduction of 15%–30%. Fig. 30 shows the energy-saving effect under short-term traffic inactivity.

6.3 Prototype performance validation for MIMO weight optimization scenarios

The DT model design is illustrated in Fig. 31. During the wireless twin design phase, various critical information such as latitude and longitude, equipment manufacturer information, engineering parameters, and simulation algorithms is bound to the digital twin. These pieces of information are the foundation for constructing and executing wireless simulation tasks. They provide necessary data support for the subsequent construction of wireless digital twin scenarios and the execution of simulation tasks.

After completing the wireless digital twin design, the current network management as well as O&M center (OMC) must be interfaced for the instantiation of digital twin. Using this instantiated information and geographic information system related technologies, the scenario required for wireless simulation is constructed. This scenario includes essential information such as network topology, equipment location, and network parameter configuration, providing an actual operating environment for subsequent simulation tasks.

1. Prototype MIMO weight optimization

The MIMO weight optimization prototype is implemented in multiple servers. The operation can be easily performed by communication engineers who work with MIMO coverage optimization, with comparison of RSRP in the area before and after optimization and the evaluation results (Fig. 32 and Table 4).

Massive MIMO antenna weight optimization based on multiagent RL (Jiang L et al., 2023) aims to

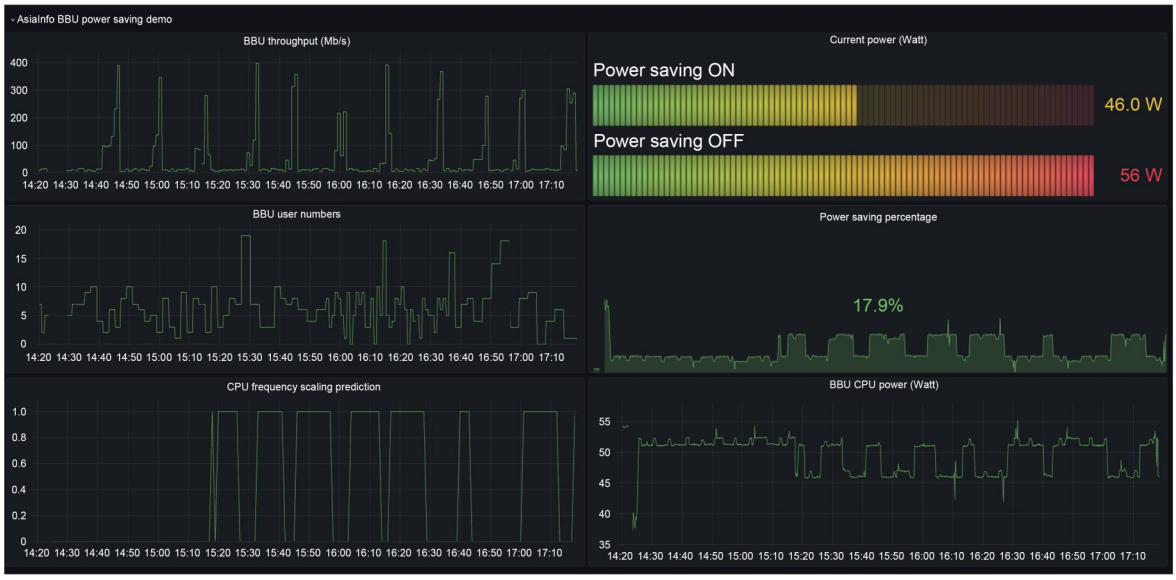


Fig. 30 Experiment selection and efficient management (BBU: baseband unit; CPU: central processing unit)

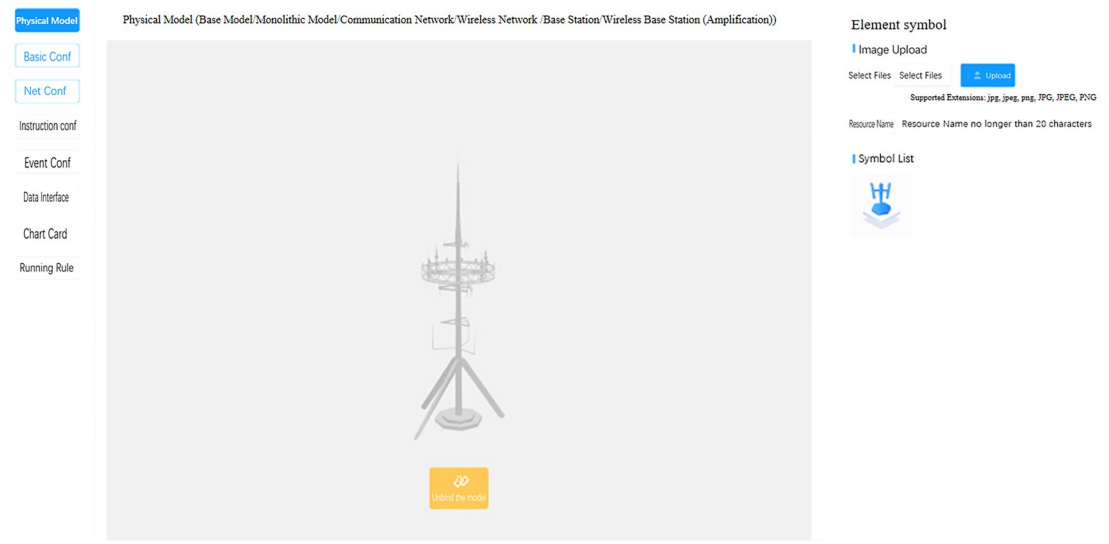


Fig. 31 Digital twin model design

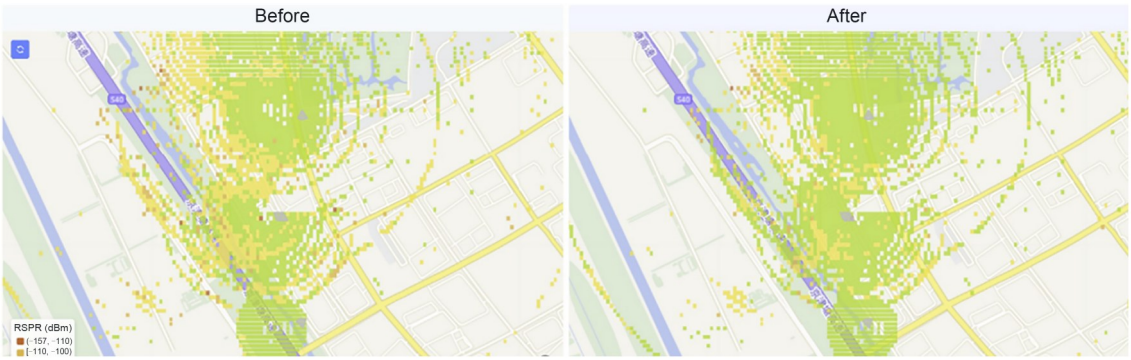


Fig. 32 Comparison of RSRP in the area before and after optimization using MIMO weight optimization prototype (MIMO: multi-input multi-output; RSRP: reference signal received power)

automatically adjust antenna parameters in real time based on the current environment. This is performed to enhance network coverage levels, avoid manpower and material consumption of manual parameter adjustment, and achieve the optimal network coverage and performance. The visualization of weight optimization is presented in Fig. 33.

Table 4 Results of regional optimization evaluation

Adjustment	Coverage rate (%)	RSRP (dBm)
Before	97.98	−87.23
After	98.74	−85.72

RSRP: reference signal received power

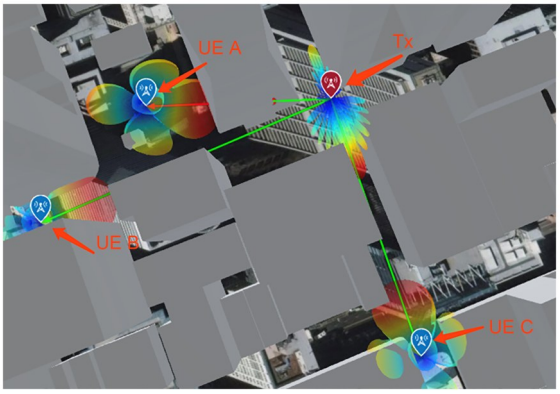


Fig. 33 Visualization of weight optimization (UE: user equipment; Tx: transmitter)

2. Verification

By employing multiagent RL algorithms combined with the network environment generated by DTNs, antenna parameters can be automatically adjusted in real time according to the current environment. Regarding AI algorithm selection, considering the high density of future network cells and the extremely complex interference situation between adjacent cells, optimization modeling must consider the impact of multiple cells simultaneously. This will allow for the selection of multiagent RL algorithms for modeling. Using training strategies in the joint action space of multiple cells, multicell collaborative optimization is achieved. Thus, weak coverage areas are optimized and no adverse effects are posed on the adjacent areas. Fig. 34 shows that multicell joint optimization can achieve a 0.5% gain in coverage rate and a 15.4% improvement in signal-to-interference-plus-noise ratio (SINR) of weak coverage areas.

7 Research challenges and open issues

7.1 Splitting and definition of services

6G service-based RANs face numerous challenges, primarily in service definition, network performance, test and maintenance, network energy efficiency, heterogeneous hardware, network security, initial costs,

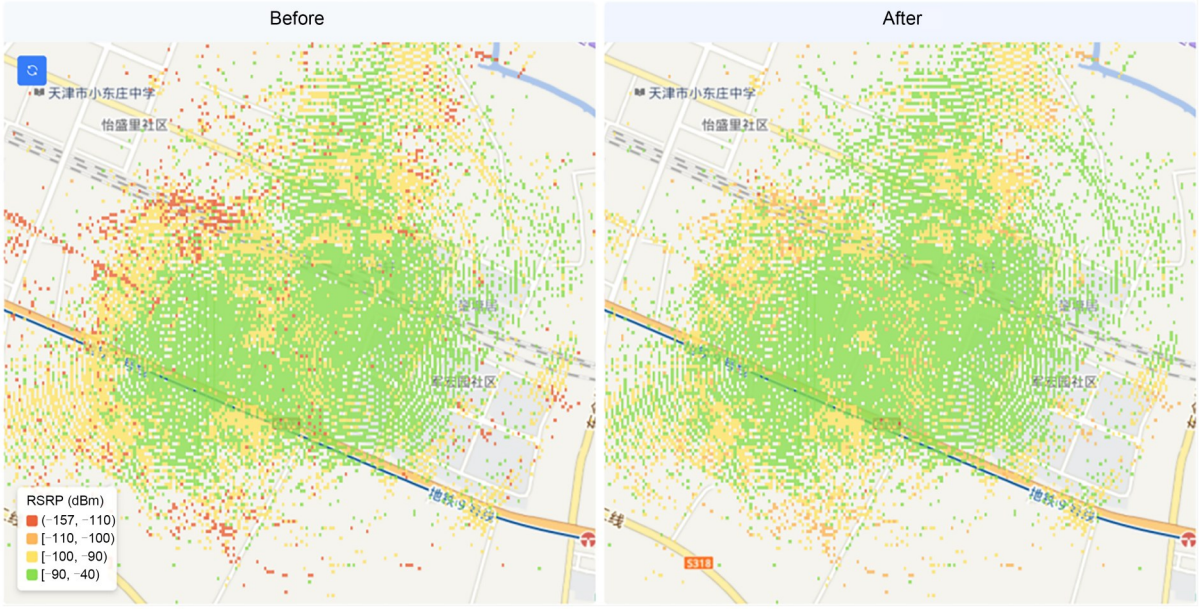


Fig. 34 Comparison of MIMO weight adjustment (left: before; right: after). RSRP: reference signal received power

and organizational structure. Among these, the splitting and definition of services are the most significant challenges encountered by current service-based RANs.

In the field of IT, where microservice exploration is highly mature, a specific and well-defined algorithm that can efficiently perform service splitting is unavailable. In the current service-based architecture of 5G core (5GC), the number of NF splits is considerably large and its design is quite complex. Addressing this issue also remains a challenge for the 6G service-based architecture. Compared with CN, RAN demands more stringent performance and service splitting will inevitably lead to a certain degree of performance loss. Minimizing performance loss is another critical issue in the splitting and definition of services. Thus, the necessity and reasonableness of RAN service splitting have become key focal points that need to be addressed. In the future, trade-offs and balances among high performance, diversity, and cost-effectiveness will have to be reassessed, rather than simply defining high performance as a system optimization goal.

7.2 NetGPT

Due to the limitation of poor generalizability in traditional small AI models, different small AI models are required for various network autonomy tasks. This increases the difficulty of orchestration and management of AI tasks within the native AI network architecture. Inspired by the superior generalizability of LLMs, a promising and meaningful solution is to train a large AI model for diverse network autonomy tasks, as illustrated in NetGPT (Jiang L et al., 2023). Using a large model as its base, an AI agent can become the best implementation form of NetGPT. As the AI agent has perception, tool calling, and self-learning abilities, it can bring a higher degree of autonomy to the network.

Despite these advantages, designing and training a large AI model in the targeted field of wireless communication is difficult. This is because network autonomy tasks differ considerably from those in natural language processing. In addition, NetGPT will pose many challenges to the transmission, storage, computation, energy consumption, and other aspects of the network. Therefore, the approach of constructing a distributed learning architecture in a wireless network with limited multidimensional resources to promote training and inference of NetGPT must be evaluated, particularly in an effectively distributed manner. This may lead to more efficient distributed learning methods, which also require continued research efforts. To enable an AI agent with the perception, tool calling, and self-learning abilities and ensure the efficient communication between different AI agents, future networks will have to be improved such as by designing new architecture and protocol for AI agents.

7.3 Efficiency and accuracy of digital twin modeling

DTNs must be modeled in different scenarios. The modeling accuracy directly determines the performance and effectiveness of the prediction, simulation, and optimization of DTNs. Fragmented network modeling based on the single-domain model library is widely used for digital twin modeling of network elements, mobile devices, and wireless channels. Moreover, systematic coarse-grained network modeling based on a multidomain model library has been proposed for this purpose. However, the accuracy, simulation efficiency, and real-time performance of these two modeling methods cannot meet the requirements of high-precision DTN modeling. Efficient and accurate modeling continues to remain a challenge, and whether high-order modeling or an LLM will facilitate this process has to be researched.

High-order models can provide finer-grained network modeling, thereby improving model accuracy and simulation efficiency. For instance, GNNs and other methods are used to accurately characterize the attributes and connection relationships of network nodes. Combined with generative AI, the global perspective and interpretability of GNNs increase based on the attention mechanism. LLMs can flexibly and efficiently meet the digital twin modeling needs in different scenarios by extracting small models.

7.4 Heterogeneous hardware selection, management, and scheduling for AI computing

The diverse deployment scales and fragmented scenarios in RANs create a substantial demand for high-performance heterogeneous hardware; however, several challenges are encountered. First, building efficient, affordable, and energy-efficient hardware platforms requires further research that involves optimization across chip design, accelerators, servers, and system architecture. Second, the management of

heterogeneous resources continues to pose significant difficulties. Due to stringent deployment conditions and varied requirements in RANs, multiple types of heterogeneous resources often coexist. Ensuring their effective operation and unified management necessitates optimization at multiple levels, including heterogeneous hardware, drivers, and cloud components. Last, fine-grained and multidimensional scheduling of AI computing resources such as GPUs has garnered increased attention. Therefore, addressing the real-time processing needs and low power consumption requirements of RANs still demands further exploration into frameworks, algorithms, and scheduling mechanisms.

8 Conclusions

To be more intelligent, agile, and elastic, the evolution of 5G toward 6G mobile networks can be further enhanced and evolved using a native cloud, native AI, and NDT. To this end, a 6G autonomous RAN was proposed herein. First, a microservice-based architecture was proposed to re-architect the protocol stack of the air interface in RANs. This enabled the flexible orchestration of services and functions on demand, as well as the customization and personalization of services in a cloud-native manner. Second, a native AI framework was developed to coordinate the necessary algorithms, data, and computing resources required by AI use cases, ensuring QoS-guaranteed AI services. Third, a DTN was established to serve as a virtual environment for AI model training, prevalidation, and performance tuning, which helps minimize the potential risks associated with deploying poorly trained AI models in network operations. The combination of native AI and NDT enhanced network autonomy by enabling closed-loop management and optimization of RANs.

Contributors

Guangyi LIU designed the research. All the authors drafted and revised the paper.

Conflict of interest

Guangyi LIU is the executive lead editor of this special issue; Jianhua ZHANG, Yang YANG, Yan ZHANG, and Jiangzhou WANG are guest editors of this special issue. Jianhua ZHANG is also an executive associate editor-in-chief of *Frontiers of Information Technology & Electronic Engineering*. They were not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Open access

References

3GPP, 2017. Study on New Radio Access Technology: Radio Access Architecture and Interfaces. TR 38.801, France.

3GPP, 2023a. Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Service Data Adaptation Protocol (SDAP) Specification. TS 37.324, France.

3GPP, 2023b. Management and Orchestration; Levels of Autonomous Network. TS 28.100, France.

3GPP, 2023c. NR; Medium Access Control (MAC) Protocol Specification. TS 38.321, France.

3GPP, 2023d. NR; Packet Data Convergence Protocol (PDCP) Specification. TS 38.323, France.

3GPP, 2023e. NR; Radio Link Control (RLC) Protocol Specification. TS 38.322, France.

3GPP, 2023f. NR; Services Provided by the Physical Layer. TS 38.202, France.

Abdullah M, Madain A, Jararweh Y, 2022. ChatGPT: fundamentals, applications and social impacts. 9th Int Conf on Social Networks Analysis, Management and Security, p.1-8. https://doi.org/10.1109/SNAMS58071.2022.10062688

Adem N, Benfaid A, Harib R, et al., 2021. How crucial is it for 6G networks to be autonomous? https://doi.org/10.48550/arXiv.2106.06949

Almasan P, Ferriol-Galmés M, Paillisse J, et al., 2022. Network digital twin: context, enabling technologies, and opportunities. *IEEE Commun Mag*, 60(11):22-27. https://doi.org/10.1109/MCOM.001.2200012

Banerjee A, Mwanje SS, Carle G, 2021. An intent-driven orchestration of cognitive autonomous networks for RAN management. 17th Int Conf on Network and Service Management, p.380-384. https://doi.org/10.23919/CNSM52442.2021.9615505

Benzaid C, Taleb T, 2020. AI-driven zero touch network and

service management in 5G and beyond: challenges and research directions. *IEEE Netw*, 34(2):186-194. https://doi.org/10.1109/MNET.001.1900252

Bhat JR, Alqahtani SA, 2021. 6G ecosystem: current status and future perspective. *IEEE Access*, 9:43134-43167. https://doi.org/10.1109/ACCESS.2021.3054833

Bonati L, 2022. Softwarized Approaches for the Open RAN of NextG Cellular Networks. PhD Dissemination, Northeastern University, Boston, USA.

Boutaba R, Shahriar N, Salahuddin MA, et al., 2021. AI-driven closed-loop automation in 5G and beyond mobile networks. Proc 4th FlexNets Workshop on Flexible Networks Artificial Intelligence Supported Network Flexibility and Agility, p.1-6. https://doi.org/10.1145/3472735.3474458

Cha J, Moon Y, Cho S, et al., 2022. RAN-CN converged user-plane for 6G cellular networks. IEEE Global Communications Conf, p.2843-2848. https://doi.org/10.1109/GLOBECOM48099.2022.10001487

Chen YX, Li RP, Zhao ZF, et al., 2024. NetGPT: an AI-native network architecture for provisioning beyond personalized generative services. *IEEE Netw*, 38(6):404-413. https://doi.org/10.1109/MNET.2024.3376419

China Mobile, 2021. China Mobile Network Autonomous Driving White Paper (in Chinese). extension://bfdogplmndidlpjfhoijckpakkdjkkil/pdf/viewer.html?file=https%3A%2F%2Fkxlabs.10086.cn%2Ffiles%2F1626350861865-520854.pdf [Accessed on July 26, 2024].

China Mobile, 2022. 6G Service-Based RAN White Paper (in Chinese). extension://bfdogplmndidlpjfhoijckpakkdjkkil/pdf/viewer.html?file=https%3A%2F%2F13115299.s21i.faiusr.com%2F61%2F1%2FABUIABA9GAAg_smAkQYooOzG3wQ.pdf [Accessed on Aug. 1, 2024].

China Mobile, 2023. 6G Service-Based RAN White Paper (in Chinese). extension://bfdogplmndidlpjfhoijckpakkdjkkil/pdf/viewer.html?file=https%3A%2F%2F13115299.s21i.faiusr.com%2F61%2F1%2FABUIABA9GAAg-be-qQYoivyeKA.pdf [Accessed on July 28, 2024].

Choi J, Sharma N, Gantha SS, et al., 2022. RAN-CN converged control-plane for 6G cellular networks. IEEE Global Communications Conf, p.1253-1258. https://doi.org/10.1109/GLOBECOM48099.2022.10001281

Coronado E, Behravesh R, Subramanya T, et al., 2022. Zero touch management: a survey of network automation solutions for 5G and 6G networks. *IEEE Commun Surv Tut*, 24(4):2535-2578. https://doi.org/10.1109/COMST.2022.3212586

Cui YP, Lv TJ, Ni W, et al., 2023. Digital twin-aided learning for managing reconfigurable intelligent surface-assisted, uplink, user-centric cell-free systems. *IEEE J Sel Areas Commun*, 41(10):3175-3190. https://doi.org/10.1109/JSAC.2023.3310050

DeAlmeida JM, Pontes CFT, Dasilva LA, et al., 2021. Abnormal behavior detection based on traffic pattern categorization in mobile networks. *IEEE Trans Netw Serv Manag*, 18(4):4213-4224. https://doi.org/10.1109/TNSM.2021.3125019

Deng J, Tian KC, Zheng QB, et al., 2022. Cloud-assisted distributed edge brains for multi-cell joint beamforming optimization for 6G. *China Commun*, 19(3):36-49. https://doi.org/10.23919/JCC.2022.03.003

Duan XY, Kang HH, Zhang JJ, 2022. Autonomous network technology innovation in digital and intelligent era. *ZTE Commun*, 20(4):52-61. https://doi.org/10.12142/ZTECOM.202204007

Eriksson D, Pearce M, Gardner JR, et al., 2019. Scalable global optimization via local Bayesian optimization. Proc 33rd Conf on Neural Information Processing Systems, p.5496-5507.

Ferriol-Galmés M, Suárez-Varela J, Paillissé J, et al., 2022. Building a digital twin for network optimization using graph neural networks. *Comput Netw*, 217:109329. https://doi.org/10.1016/j.comnet.2022.109329

Gill SS, Xu MX, Ottaviani C, et al., 2022. AI for next generation computing: emerging trends and future directions. *Int Things*, 19:100514. https://doi.org/10.1016/j.iot.2022.100514

Hazra A, Morichetta A, Murturi I, et al., 2024. Distributed AI in zero-touch provisioning for edge networks: challenges and research directions. *Computer*, 57(3):69-78. https://doi.org/10.1109/MC.2023.3334913

He WL, Zhang C, Deng J, et al., 2023. Conditional generative adversarial network aided digital twin network modeling for massive MIMO optimization. IEEE Wireless Communications and Networking Conf, p.1-5. https://doi.org/10.1109/WCNC55385.2023.10118756

He XW, Yang ZM, Xiang Y, et al., 2023. NWDAF in 3GPP 5G advanced: a survey. 3rd Int Conf on Electronic Information Engineering and Computer Science, p.756-761. https://doi.org/10.1109/EIECS59936.2023.10435433

Hu F, Hao Q, Bao K, 2014. A survey on software-defined network and OpenFlow: from concept to implementation. *IEEE Commun Surv Tut*, 16(4):2181-2206. https://doi.org/10.1109/COMST.2014.2326417

Huawei, 2023. Autonomous Driving Network (ADN). https://carrier.huawei.com/en/adn [Accessed on July 23, 2024].

Hui SD, Wang HD, Li T, et al., 2023. Large-scale urban cellular traffic generation via knowledge-enhanced GANs with multi-periodic patterns. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.4195-4206. https://doi.org/10.1145/3580305.3599853

Institute CMCCR, 2022. 6G Autonomous Mobile Network Enabled by Digital Twin Network White Paper (in Chinese). https://www.sgpjbg.com/baogao/64570.html [Accessed on July 30, 2024].

Ismail T, Mahmoud HHM, 2020. Optimum functional splits for optimizing energy consumption in V-RAN. *IEEE Access*, 8:194333-194341. https://doi.org/10.1109/ACCESS.2020.3033879

ITU-R, 2023. Framework and Overall Objectives of the Future

Development of IMT for 2030 and Beyond. https://techblog.comsoc.org/2023/01/29/ [Accessed on Aug. 12, 2024].

Jain R, Paul S, 2013. Network virtualization and software defined networking for cloud computing: a survey. *IEEE Commun Mag*, 51(11):24-31. https://doi.org/10.1109/MCOM.2013.6658648

Jiang L, Wang XS, Yang AD, et al., 2023. An efficient multi-agent optimization approach for coordinated massive MIMO beamforming. IEEE Int Conf on Communications, p.5632-5638. https://doi.org/10.1109/ICC45041.2023.10279724

Jiang W, Han B, Habibi MA, et al., 2021. The road towards 6G: a comprehensive survey. *IEEE Open J Commun Soc*, 2:334-366. https://doi.org/10.1109/OJCOMS.2021.3057679

Kalogiros C, Muschamp P, Caruso G, et al., 2021. Capabilities of business and operational support systems for pre-commercial 5G testbeds. *IEEE Commun Mag*, 59(12):58-64. https://doi.org/10.1109/MCOM.003.2001059

Kamran R, Kiran S, Jha P, et al., 2024. Green 6G: energy awareness in design. 16th Int Conf on Communication Systems & Networks, p.1122-1125. https://doi.org/10.1109/COMSNETS59351.2024.10427334

Kaur J, Khan MA, 2022. Sixth generation (6G) wireless technology: an overview, vision, challenges and use cases. IEEE Region 10 Symp, p.1-6. https://doi.org/10.1109/TENSYMP54529.2022.9864388

Khan TA, Abbas K, Muhammad A, et al., 2022. An intent-driven closed-loop platform for 5G network service orchestration. *Comput Mater Con*, 70(3):4323-4340. https://doi.org/10.32604/cmc.2022.017118

Kim H, Feamster N, 2013. Improving network management with software defined networking. *IEEE Commun Mag*, 51(2):114-119. https://doi.org/10.1109/MCOM.2013.6461195

Lähdekorpi P, Hronec M, Jolma P, et al., 2017. Energy efficiency of 5G mobile networks with base station sleep modes. IEEE Conf on Standards for Communications and Networking, p.163-168. https://doi.org/10.1109/CSCN.2017.8088616

Li LL, 2024. A survey on intelligence-endogenous network: architecture and technologies for future 6G. *Intell Conv Netw*, 5(1):53-67. https://doi.org/10.23919/ICN.2024.0005

Li N, Liu GY, Zhang HM, et al., 2022a. Micro-service-based radio access network. *China Commun*, 19(3):1-15. https://doi.org/10.23919/JCC.2022.03.001

Li N, Liu GY, Zhang HM, et al., 2022b. Service-based RAN: the next phase of cloud RAN. IEEE Globecom Workshops, p.1206-1211. https://doi.org/10.1109/GCWkshps56602.2022.10008666

Li Q, Ding ZR, Tong XP, et al., 2022. 6G cloud-native system: vision, challenges, architecture framework and enabling technologies. *IEEE Access*, 10:96602-96625. https://doi.org/10.1109/ACCESS.2022.3205341

Liu GY, Jin J, Wang QX, 2020a. Vision and requirements of 6G: digital twin and ubiquitous intelligence. *Mob Commun*, 44(6):3-9 (in Chinese). https://doi.org/10.3969/j.issn.1006-1010.2020.06.001

Liu GY, Huang YH, Li N, et al., 2020b. Vision, requirements and network architecture of 6G mobile network beyond 2030. *China Commun*, 17(9):92-104. https://doi.org/10.23919/JCC.2020.09.008

Liu GY, Li N, Deng J, et al., 2022. The SOLIDS 6G mobile network architecture: driving forces, features, and functional topology. *Engineering*, 8:42-59. https://doi.org/10.1016/j.eng.2021.07.013

Liu GY, Zhang HM, Tong Z, et al., 2024. 6G mobile information network architecture: migrate from communication to XaaS. *Sci Sin Inform*, 54(5):1236-1266 (in Chinese). https://doi.org/10.1360/SSI-2023-0339

Liu ZH, Zhang M, Zhang CH, et al., 2023. 6G network self-evolution: generating core networks. IEEE Int Conf on Communications Workshops, p.625-630. https://doi.org/10.1109/ICCWorkshops57953.2023.10283790

Long QY, Chen YL, Zhang HJ, et al., 2022. Software defined 5G and 6G networks: a survey. *Mob Netw Appl*, 27(5):1792-1812. https://doi.org/10.1007/s11036-019-01397-2

Lu YL, Maharjan S, Zhang Y, 2021. Adaptive edge association for wireless digital twin networks in 6G. *IEEE Int Things J*, 8(22):16219-16230. https://doi.org/10.1109/JIOT.2021.3098508

Maharana K, Mondal S, Nemade B, 2022. A review: data pre-processing and data augmentation techniques. *Glob Trans Proc*, 3(1):91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Mahbub M, Shubair RM, 2022. Energy efficient maximization of user association employing IRS in mmWave multi-tier 6G networks. IEEE Int Conf on Sensing, Communication, and Networking, p.25-30. https://doi.org/10.1109/SECONWorkshops56311.2022.9926334

Mai VS, La RJ, Zhang T, et al., 2022. End-to-end quality-of-service assurance with autonomous systems: 5G/6G case study. IEEE 19th Annual Consumer Communications & Networking Conf, p.644-651. https://doi.org/10.1109/CCNC49033.2022.9700514

Mao BM, Tang FX, Kawamoto Y, et al., 2022. AI models for green communications towards 6G. *IEEE Commun Surv Tut*, 24(1):210-247. https://doi.org/10.1109/COMST.2021.3130901

Mehmood K, Kralevska K, Palma D, 2023. Intent-driven autonomous network and service management in future cellular networks: a structured literature review. *Comput Netw*, 220:109477. https://doi.org/10.1016/j.comnet.2022.109477

Nidhi, Mihovska A, Kumar A, et al., 2022. Business opportunities for beyond 5G and 6G networks. 25th Int Symp on Wireless Personal Multimedia Communications, p.543-548. https://doi.org/10.1109/WPMC55625.2022.10014752

Niemöller J, Müller E, Maggiari M, et al., 2024. Evolving

service management towards intent-driven autonomous networks. *Ericss Technol Rev*, 2024(3):2-7.

Niknam S, Dhillon HS, Reed JH, 2020. Federated learning for wireless communications: motivation, opportunities, and challenges. *IEEE Commun Mag*, 58(6):46-51. https://doi.org/10.1109/MCOM.001.1900461

Patwardhan N, Marrone S, Sansone C, 2023. Transformers in the real world: a survey on NLP applications. *Information*, 14(4):242. https://doi.org/10.3390/info14040242

Pivoto DGS, Rezende TT, Facina MSP, et al., 2023. A detailed relevance analysis of enabling technologies for 6G architectures. *IEEE Access*, 11:89644-89684. https://doi.org/10.1109/ACCESS.2023.3301811

Qin Z, Deng SG, Yan XQ, et al., 2023. 6G data plane: a novel architecture enabling data collaboration with arbitrary topology. *Mob Netw Appl*, 28(1):394-405. https://doi.org/10.1007/s11036-023-02093-y

Raj DRR, Shaik TA, Hirwe A, et al., 2023. Building a digital twin network of SDN using knowledge graphs. *IEEE Access*, 11:63092-63106. https://doi.org/10.1109/ACCESS.2023.3288813

Rohani R, 2023. Function vs Service vs Platform. https://rlohani.medium.com/function-vs-service-vs-platform-e2ac25445167 [Accessed on July 29, 2024].

Shahjalal M, Kim W, Khalid W, et al., 2023. Enabling technologies for AI empowered 6G massive radio access networks. *ICT Exp*, 9(3):341-355. https://doi.org/10.1016/j.icte.2022.07.002

Sun YT, Zhang JH, Yu L, et al., 2023. How to define the propagation environment semantics and its application in scatterer-based beam prediction. *IEEE Wirel Commun Lett*, 12(4):649-653. https://doi.org/10.1109/LWC.2023.3237827

Tang QQ, Xie RC, Fang ZR, et al., 2024a. Joint service deployment and task scheduling for satellite edge computing: a two-timescale hierarchical approach. *IEEE J Sel Areas Commun*, 42(5):1063-1079. https://doi.org/10.1109/JSAC.2024.3365889

Tang QQ, Xie RC, Feng L, et al., 2024b. SIaTS: a service intent-aware task scheduling framework for computing power networks. *IEEE Netw*, 38(4):233-240. https://doi.org/10.1109/MNET.2023.3326239

Tao ZY, Xu W, You XH, 2023. Digital twin assisted deep reinforcement learning for online admission control in sliced network. https://doi.org/10.48550/arXiv.2310.09299

TG3, 2023. Wireless Network Data Dictionary White Paper (in Chinese). https://www.6g-ana.com/upload/file/20231214/6383817255076725588362734.pdf [Accessed on Aug. 16, 2024].

Umoga UJ, Sodiya EO, Ugwuanyi ED, et al., 2024. Exploring the potential of AI-driven optimization in enhancing network performance and efficiency. *Magna Sci Adv Res Rev*, 10(1):368-378. https://doi.org/10.30574/msarr.2024.10.1.0028

Villalobos P, Ho A, Sevilla J, et al., 2024. Will we run out of data? Limits of LLM scaling based on human-generated data. https://doi.org/10.48550/arXiv.2211.04325

Wang S, Sun T, Yang HW, et al., 2020. 6G network: towards a distributed and autonomous system. 2nd 6G Wireless Summit, p.1-5. https://doi.org/10.1109/6GSUMMIT49458.2020.9083888

Wang SF, Chen HM, Ouyang Y, et al., 2023a. Digital twin network application requirement on green coordination of computing and networking. IEEE 3rd Int Conf on Digital Twins and Parallel Intelligence, p.1-6. https://doi.org/10.1109/DTPI59677.2023.10365446

Wang SF, Chen HM, Ouyang Y, et al., 2023b. Elastic digital twin network modeling fulfilling twining dynamic in network life cycle. IEEE 3rd Int Conf on Digital Twins and Parallel Intelligence, p.1-7. https://doi.org/10.1109/DTPI59677.2023.10365450

Wu JJ, Li RP, An XL, et al., 2021. Toward native artificial intelligence in 6G networks: system design, architectures, and paradigms. https://doi.org/10.48550/arXiv.2103.02823

Yan XQ, An XL, Yu WX, et al., 2021. A blockchain-based subscriber data management scheme for 6G mobile communication system. IEEE Globecom Workshop, p.1-6. https://doi.org/10.1109/GCWkshps52748.2021.9682154

Yang CG, Mi XR, Ouyang Y, et al., 2023. Smart intent-driven network management. *IEEE Commun Mag*, 61(1):106-112. https://doi.org/10.1109/MCOM.002.2200119

Yang Y, Ma ML, Wu HQ, et al., 2023. 6G network AI architecture for everyone-centric customized services. *IEEE Netw*, 37(5):71-80. https://doi.org/10.1109/MNET.124.2200241

Yang YQ, Yang SS, Zhao C, et al., 2024. TelOps: AI-driven operations and maintenance for telecommunication networks. *IEEE Commun Mag*, 62(4):104-110. https://doi.org/10.1109/MCOM.003.2300055

Yaqoob M, Trestian R, Tatipamula M, et al., 2024. Digital-twin-driven end-to-end network slicing toward 6G. *IEEE Int Comput*, 28(2):47-55. https://doi.org/10.1109/MIC.2023.3332252

Younes M, Louet Y, 2022. Joint optimization of energy consumption and spectral efficiency for 5G/6G point-to-point networks. 3rd URSI Atlantic and Asia Pacific Radio Science Meeting, p.1-4. https://doi.org/10.23919/AT-AP-RASC54737.2022.9814348

Yu L, Zhang YX, Zhang JH, et al., 2022. Implementation framework and validation of cluster-nuclei based channel model using environmental mapping for 6G communication systems. *China Commun*, 19(4):1-13. https://doi.org/10.23919/JCC.2022.04.001

Zhang D, Zhao YJ, Zhao ZC, et al., 2024. Research on intelligent operation architecture and evolution of 6G network. *Des Technol Post Telecommun*, 2024(3):32-37 (in Chinese). https://doi.org/10.12045/j.issn.1007-3043.2024.03.007

Zhang LF, Hu ZY, Li YZ, et al., 2022. Architecture and applications of wireless autonomous network. IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles,

p.2048-2051. https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00296

Zhang SY, Li T, Hui SD, et al., 2023. Deep transfer learning for city-scale cellular traffic generation through urban knowledge graph. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.4842-4851. https://doi.org/10.1145/3580305.3599801

Zhao BR, Cui QM, Liang SY, et al., 2022. Green concerns in federated learning over 6G. *China Commun*, 19(3):50-69. https://doi.org/10.23919/JCC.2022.03.004

Zhu YH, Chen DY, Zhou C, et al., 2021. A knowledge graph based construction method for digital twin network. IEEE 1st Int Conf on Digital Twins and Parallel Intelligence, p.362-365. https://doi.org/10.1109/DTPI52967.2021.9540177

Ziegler V, Viswanathan H, Flinck H, et al., 2020. 6G architecture to connect the worlds. *IEEE Access*, 8:173508-173520. https://doi.org/10.1109/ACCESS.2020.3025032

Zong JY, Liu HT, Liu Y, et al., 2022. Service-based architecture evolution of radio access network towards 6G. Proc 12th Int Conf on Computer Engineering and Networks, p.525-534. https://doi.org/10.1007/978-981-19-6901-0_56