GTI Network Computing Enabling Terminal and Network Coordination White Paper

https://gtigroup.org/

GTI Network Computing Enabling Terminal and Network Coordination White Paper

Version:	V1.0
Deliverable Type	□ **Procedural Document** ☑ **Working Document**
Confidential Level	☑ **Open to GTI Operator Members** ☑ **Open to GTI Partners** □ **Open to Public**
Program	Network WG
Working Group	N/A
Project	Project 1: Network intelligence Project 2: Application intelligence
Task	N/A
Source members	China Mobile, Intel, OPPO, vivo, TCL, Huawei, AsiaInfo, Rokid
Support members	
Editor	China Mobile: Yushuang Hu, Zhenglei Huang, Huiqing Sun, Zehao Chen Intel: Yanqing Chen OPPO: Yang Xu, Changhong Shan, Yang Shen, Jingran Chen, Yaxin Wang, Cong Shi vivo: Yannan Yuan, Yanchao Kang, Dajie Jiang, Xiaobo Wu TCL: Xiaoyong Tang, Yincheng Zhang

	Huawei: Lei Wang, Quan Zhu, Kan Zeng, Xiaochun Zheng AsiaInfo: Ye Ouyang, Shoufeng Wang, Jianchao Guo, Leiming Ma
Last Edit Date	01-03-2025
Approval Date	01-03-2025

Document History

Date	Meeting #	Version #	Revision Contents
01-03-2025		V1.0	Initial Draft

Table of Contents

1 Executive Summary

To build the capability for intelligent terminal-network coordination at the beginning of 6G, it is necessary for the telecommunications industry, including operators, terminal manufacturers, OTT application providers, and communication equipment vendors, to work together to promote industrial reform and build the core competitive edge of the future telecom network technology market.

2 Abbreviations

Abbreviation	**Explanation**
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
ATW	Asynchronous Timewarp
FaaS	Function as a Service
LLM	Large Language Models
MEC	Multi-access Edge Computing
MTP	Motion-to-Photon
NAS	Non-Access-Stratum
NPL	Natural language processing
OTT	Over-The-Top
PCC	Private Compute Clouds
TNC	Terminal-Network Coordination
TTS	Text-to-Speech
XR	Extended Reality

3 Introduction

3.1 The Evolution of Future Network

Future networks face significantly increased demands due to the rapid advancement of Artificial Intelligence (AI), particularly with the rise of multi-modal Large Language Models (LLM), alongside the growing use of immersive technologies like extended Reality (XR). These demands will far exceed the traditional network metrics of low latency, high bandwidth, and wide coverage. In this context, the future network will no longer be limited to serving as a mere connectivity infrastructure. Instead, it will enable real-time, in-depth interactions not only between humans and AI agents but also among AI agents themselves. This requires the network to break through the limitations of terminal types, information modalities, and interaction scenarios, achieving intelligent coordination across domains, devices, and platforms. Looking ahead, network evolution will no longer solely focus on providing basic connectivity. Rather, it will evolve into an intelligent ecosystem that integrates computing, sensing, and other services. Consequently, future networks will not only meet foundational connection needs but will also offer capabilities such as computing and sensing that go “beyond connectivity”.

The evolution of 6G networks is driving a surge in the development of emerging scenarios such as immersive communication, AI agents, and integrated sensing. Under this trend, users' desire for high-quality service experiences is growing stronger, and their demands for service levels are correspondingly increasing. At the same time, the future world is moving towards a direction where distributed multi-dimensional computing capability, ubiquitous holographic sensing, and the parallel development of high-intelligence and high-computing-capability applications are progressing together. This transformative shift imposes new requirements on the terminal-network systems within 6G, demanding comprehensive capabilities for sensing information gathering, robust computing services, precise intelligent decision-making, and efficient execution, among others. This represents an unprecedented challenge to the collaborative working capabilities of terminal devices and networks.

To address these challenges, **6G networks will leverage the computing capability of the network to empower terminals through a transformative Terminal-Network Coordination mechanism. In the future Terminal- Network Coordination mechanism**, the network will no longer be limited to providing basic connectivity. Instead, it will become a key pillar supporting terminals in achieving enhanced computing capability, greater storage capacity, and longer battery life. Furthermore, the network will play a key role in efficiently managing and scheduling computing tasks to improve service quality. Specifically, the network must break traditional service boundaries based on the Terminal-Network Coordination mechanism. It will provide native services such as computing and sensing to extend terminal capabilities. Achieving this requires deeper integration between the network and terminals, through more efficient service-oriented mechanisms, intelligent communication negotiation, service management, and advanced computing service frameworks. By understanding user needs, the network will provide multi-layered services in an intelligent manner, empowering terminals to adapt more flexibly to complex application scenarios.

Therefore, the core of future networks will no longer be the mere provision of "connectivity." Instead, fueled by network computing, it will focus on intelligent Terminal-Network Coordination, driving deep cooperation and symbiotic development between users, terminal devices, and networks.

3.2 What is Terminal-Network Coordination in Future Networks?

The motivation behind the Terminal-Network Coordination mechanism stems from the fact that, as terminal services, capabilities, and forms continue to evolve, a gap has emerged between the capabilities of terminals and their service requirements. This gap makes it increasingly difficult to effectively meet user experience expectations. For example, in terms of computing, future terminals will handle more AI-driven, computing intensive tasks, which require significant computing capability and energy consumption. This presents a serious challenge to the battery life of terminals. However, the trend toward more lightweight terminals makes frequent recharging impractical. Additionally, the current computing capabilities of terminals are insufficient to support the growing demands of services. To address this issue, terminals need to leverage the network. Through the Terminal-Network Coordination model, network-native services, such as computing service, can be integrated into the terminals. This coordination enables terminals to meet the increasingly complex AI service demands without sacrificing portability or battery life, thereby achieving the core goal of empowering terminals through the network.

Terminal-Network Coordination in future networks is a mechanism aimed at addressing the gap between the diverse service demands of terminals and the limitations of their hardware capabilities. It leverages mobile networks, utilizing their native advantages of low latency, high bandwidth, and wide-area coverage, to provide multi-dimensional native network services, such as connectivity, intelligent computing, and sensing, specifically designed for terminals. This coordination will be facilitated through standardized service interfaces between terminals and networks, enabling the negotiation of service resources and the intelligent management and scheduling of various network-side services.

The goal of the Terminal-Network Coordination mechanism is to provide direct network services to terminals, enabling their native service capabilities and breaking the boundaries of traditional network services. This will ultimately empower terminal applications comprehensively and lay a solid foundation for the diversification and prosperity of application ecosystems. The core value of Terminal-Network Coordination in future networks is to allow all manufacturers and all types of mobile terminals to use low-latency, high-bandwidth, and wide-coverage network services—such as computing capability, connectivity, and sensing—provided by mobile network in an open, standardized, and cross-vendor manner. This will enable various mobile terminals to better meet the diverse demands of different application scenarios.

4 Use Cases

4.1 AI Inference through Terminal-Network Coordination

4.1.1 The Value of Applying Network Computing Service in AI Inference

Terminals face significant challenges in running AI inference tasks such as natural language

processing (NLP) and computer vision due to strict limitations in their capabilities (computing, storage, battery). The smaller parameter models running on terminals are often limited to fewer than 10 billion parameters [1]. This further underscores the challenge faced by terminals, where limited battery capacity makes it difficult to support power-intensive computations, resulting in significant battery life constraints.

To overcome this limitation, terminals are increasingly offloading AI inference tasks to remote, more powerful computing nodes to meet the growing demand for complex tasks. Both terminal and chip industries have proposed collaborative computation solutions between terminals and remote nodes. For instance, the 'hybrid AI' concept is introduced, where AI workloads are distributed between terminals and the cloud based on context and timing [1]. Private Compute Clouds (PCC) is introduced to handle compute-intensive requests, specifically designed to process complex data [2]. In academia, research on AI task offloading mainly focuses on the decomposition of AI computation tasks on the terminal and the optimization of offloading strategies. Researchers typically base their task-splitting decisions on factors such as data characteristics and the logical structure of task workflows. In recent years, the decomposition of computational tasks has been further refined to the level of neural network layers, and even to individual neural network components as the smallest computational unit.

In this context, the expected value of applying Terminal-Network Coordination in AI inference can be summarized as follows:

- **Enhanced Terminal Computing Capability and Improved AI Accuracy:** Offloading computationally intensive AI inference tasks to more powerful remote computing nodes via the network not only addresses the latency challenges associated with large models but also compensates for the accuracy limitations of smaller local models. This approach enhances the overall accuracy of AI inference.
- **High-Quality Network Connectivity to Guarantee User Experience:** High-quality network connectivity, with low latency, high bandwidth, and wide coverage, ensures seamless AI inference, stable service operation, and efficient connectivity between terminals and network service, then ultimately enhancing the user experience.
- **Expanding the AI Application Ecosystem:** The ability to run more AI-driven, compute-intensive applications, such as large AI models, on terminals enriches the terminal ecosystem and broadens its application scope.

4.1.2 Use Case of AI Inference with Terminal-Network Coordination

In the 6G era, personal mobile assistants efficiently running multimodal AI-generated inference tasks is a typical use case for Terminal-Network Coordination. In this scenario, the user issues a translation command via voice on their mobile phone, and with the coordination between the network and the phone, the translation is completed and the translated speech is returned to the user.

1) Scenario Description:

The entire process consists of three main steps: speech-to-text conversion, LLM translation, and text-to-speech conversion:

- Speech-to-text conversion involves user-sensitive voice data. As the computing capability demands of this step are relatively low, it can be completed on the user's local phone, ensuring the protection of user privacy.
- Accurate translation models and high-performance text-to-speech conversion models (such

as those simulating different genders and tones) require substantial computing capability. If these tasks are processed locally on the terminal, the delays in the computing process would be significant. Alternatively, while compressing the models could reduce latency, it would compromise translation accuracy. Therefore, the terminal can offload the data to a remote computing node with greater computing capability for processing.

As illustrated in Figure 1, the user first speaks the Chinese sentence "今天天气怎样" ("How's the weather today?"). The phone uses the Automatic Speech Recognition (ASR) model deployed on the phone to convert the speech into text. The phone then sends the text to a remote computing node. These nodes may refer to computing resources within the operator's network or third-party resources such as cloud servers. At the remote site, an LLM is used to translate the text from Chinese to English ("How's the weather today"), while a text-to-speech (TTS) model converts the English text into speech. Finally, the TTS generates the translated speech and sends it back to the phone. As a result, the user can hear the translated English speech directly from the phone.

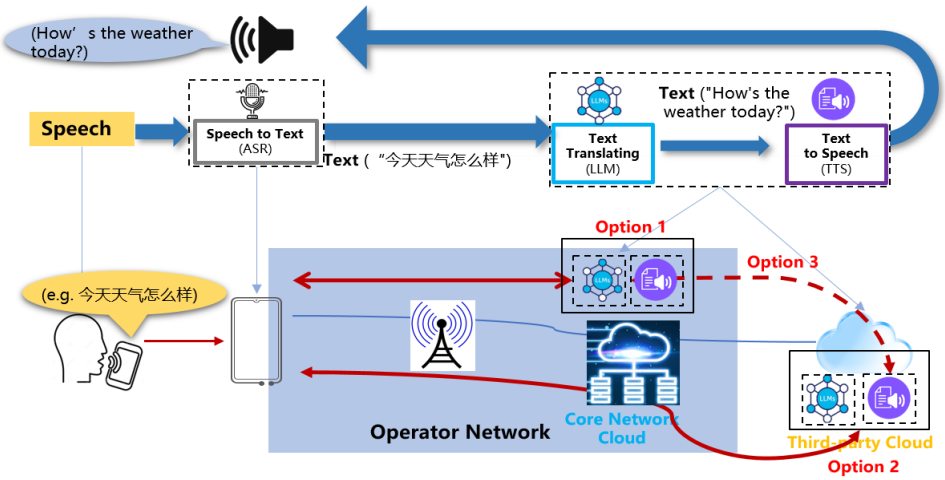

Figure 1 Use Case of AI Inference with TNC

2) Potential Solutions for Applying Terminal-Network Coordination:

The solutions for Terminal-Network Coordination primarily involve selecting remote computing services, which include base stations, core networks, edge computing, and third-party cloud computing. Due to the limitations in computing capability and cost constraints, base stations do not provide computing services in this scenario, as they rely on core network functionalities for processing computing requests. Mobile devices typically connect to computing services provided by the operator's core network, edge computing, or third-party cloud services. The specific solution options are as follows:

Option 1: Operator-Provided Computing Services: Given the real-time requirements of voice translation conversations, selecting operator network-based computing services close to the user can effectively meet latency demands. In this case, both LLM and TTS services are executed within the operator's network. The core network will aggregate terminal user requirements, network computing resources, and network conditions (such as bandwidth, latency, and packet loss rates) to manage the best computing service and routing path for the user.

Option 2: Third-Party Cloud-Provided Computing Services: In this case, both the LLM and TTS services are deployed in the cloud. The operator's network, based on user requirements and network conditions, ensures end-to-end connectivity between the terminal and third-party cloud computing services.

Option 3: Shared Computing Tasks Between Operator and Third-Party Cloud: The computing tasks are split into two parts. The operator's network provides the LLM translation service, while

the TTS model is deployed on third-party cloud computing platforms. The mobile network must not only provide computing services but also ensure high-quality connectivity to third-party cloud computing services.

4.1.3 Challenges

The main challenges affecting user experience in AI inference are how to improve the quality of inference results while minimizing response latency:

- **Low-latency inference response:**
  A. Reducing computation latency: The First Token Generation Time (TTFT) is an important metric for measuring the latency of AI inference. For example, when a mobile phone (Snapdragon 8 Gen 3) runs a lightweight, quantized LLaMa3.2-3B model, the TTFT takes at least 700ms [3]. In contrast, the TTFT of the same LLaMa2-70B model on a server (4×NVIDIA A100) is only 154ms [4]. This indicates that powerful computing support is key to reducing computation latency. Therefore, one of the core goals of Terminal-Network Coordination is for the terminal to leverage more powerful computing services from the network to reduce computation latency.
  B. Reducing transmission latency: Transmission latency from the UE to the core network or backbone network's edge router is typically under 100ms, while transmission to third-party OTT clouds incurs even higher delays. Compared to the 154ms computation latency for running the LLaMa2-70B model, the transmission latency still accounts for a significant portion of the total delay. Therefore, Terminal-Network Coordination needs to optimize the transmission path based on the user requirements, network conditions, while implementing comprehensive QoS guarantees to reduce transmission latency.
- **High-accuracy inference results:** Accuracy is a key indicator of AI inference quality. Studies show that the inference accuracy of LLM architectures increases with the size of the model's parameters [6]. However, large-parameter models require substantial computing resources to meet low-latency demands. Therefore, network needs to provide terminals with more powerful and suitable computing services to support high-precision, heavyweight model inference.

Requirements for Terminal-Network Coordination:
In summary, to address the challenges in AI inference scenarios, the network needs to ensure low-latency and high-computing capability for large-model inference tasks. This requires the Terminal-Network Coordination mechanism to consider multiple factors such as the terminal, network, and user, and to provide efficient service management and scheduling. On the one hand, the network should select the optimal transmission path and implement QoS guarantees to provide low-latency transmission services to the terminal. On the other hand, the network must possess strong native computing capabilities and efficiently allocate and manage these resources to meet the computing demands of AI inference tasks.

4.2 XR Applications through Terminal-Network Coordination

4.2.1 The Value of Applying Network Computing Service in XR Applications

Currently, the main challenge for XR glasses lies in the contradiction between the demand for

lightweight wearability and the need for sufficient battery life. On one hand, users desire a lightweight wearing experience, which makes frequent charging and the installation of heavy batteries impractical. On the other hand, the operation of XR glasses inevitably consumes energy, reducing battery life. Moreover, as the computational density and energy consumption of XR glasses increase, they may generate significant heat, posing a severe challenge for devices that are worn close to the body.

To address these challenges, XR glasses can offload some computing tasks to external computing nodes, thereby reducing their computing burden and mitigating issues related to energy consumption and heat generation while maintaining service quality. For example, to meet the needs of AR applications, the Snapdragon AR2 platform adopts a distributed processing architecture that distributes computational tasks from the glasses to a smartphone or laptop [5]. Similarly, Meta has introduced a prototype of a split-style AR glasses called Orion, which consists of three components: the glasses, a wristband that controls the glasses, and a wireless compute puck. The compute puck handles most of the computing, including graphics rendering, AI operations, and machine perception [6]. Currently, many XR glasses rely on external devices to share computational tasks, which, although making the glasses lighter, also increases the burden of carrying additional equipment for users.

Another significant characteristic of XR services is their stringent latency requirements. Many XR functions are vision-related and require an immersive experience. If latency is too high, it can lead to issues such as a synchrony between the real scene and virtual images, lag, and jitters. For Motion-to-Photon (MTP) latency, the general target is around 20 ms, and for AR applications that require real-time interaction, this requirement is even stricter. If latency exceeds the acceptable threshold, users may experience dizziness, which severely degrades the user experience. Therefore, when offloading vision-related XR tasks to network nodes, it is essential to meet strict latency requirements and find offloading nodes closer to the user than the cloud to ensure performance.

Based on the challenges and characteristics of XR services mentioned above, the value of applying terminal-network coordination mechanisms in XR services is mainly reflected in the following aspects:

- **Enhancing terminal computing capability to ensure high-quality XR experiences:** XR services involve computationally intensive tasks such as high-resolution video generation, 6-DoF volumetric video encoding and decoding, real-time rendering, etc. These tasks exceed the current computing capabilities of XR glasses and require reliance on the network to provide sufficient computing resources and services to ensure a high-quality XR experience.
- **Meeting the "low-latency, high-bandwidth" requirements of XR services:** Terminal-network coordination provides low-latency transmission, and leverages the network's native high-performance computing services to reduce computational delays, thereby meeting the extremely low-latency requirements of XR services. Additionally, for the transmission of large amounts of data such as 360° videos and high-resolution videos, terminal-network coordination can provide high-bandwidth support, fundamentally enhancing the user experience.
- **Promoting the integration of XR terminals:** Many current XR glasses rely on external devices for additional computing and wireless communication, which can be inconvenient for users. By leveraging terminal-network coordination mechanisms, this dependence on external devices can be minimized, driving the development of XR devices toward greater lightness and integration.

4.2.2 Use Case of XR Applications with Terminal-Network Coordination

1) Scenario Description:

Figure 2 illustrates a scenario where a user wears XR glasses for a day trip. In this scenario, the XR glasses provide various functions, including a shopping assistant, 3D tour guide for tourist attractions, and airport navigation. For example, while shopping, the XR glasses can display real-time prompts for points of interest such as cafe and shopping malls along the route. When passing by landmarks like the Bird's Nest, the glasses offer 3D model. Users can control the rotation, zooming in, or zooming out of the model through gesture control to obtain more details. When heading to the airport, the XR glasses need to rely on indoor positioning to provide navigation and signage services, helping users quickly locate their boarding gates.

These XR scenarios place extremely high demands on network and computing capabilities. First, the network needs to transmit high-frame-rate, high-resolution video streams in real-time, such as the 3D model of the Bird's Nest, which requires substantial network bandwidth. Second, the relevant computing tasks (such as gesture recognition, SLAM, and 3D rendering) are computationally intensive which beyond the capabilities of XR terminal devices. Therefore, the network must provide sufficient bandwidth and computing services to support these tasks. Moreover, these highly interactive, high-frame-rate XR tasks have strict real-time requirements and are extremely sensitive to latency. For example, when the user is moving quickly, the products and navigation signs along the route need to be updated in real-time to match the changing scene. Similarly, the 3D model's gesture control requires the model rendering to keep pace with the user's gestures. Any increase in latency may result in the image failing to accurately follow the gestures, leading to stuttering or other problems.

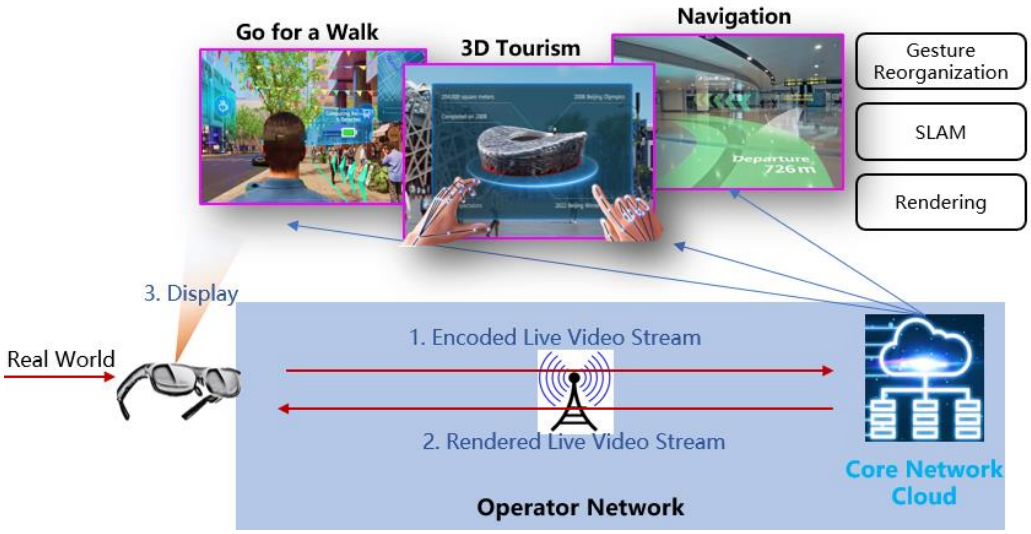

Figure 2 Use Case of XR Applications with TNC

2) Potential Solutions for Applying Terminal-Network Coordination:

For tasks with high computing capability demands, high latency sensitivity, and large data transfers, deploying computing nodes closer to users (such as in the operator's core network or edge computing nodes) is more effective in reducing latency and saving bandwidth compared to using distant third-party cloud nodes. Take Solution Option1 in Section 2.1.2 as an example:

The XR glasses perform local privacy protection and encoding of the video stream. Subsequently, the computationally intensive tasks are sent via the wireless network to a computing node within

the mobile core network that is closer to the user. The core network computing node executes computational services such as image recognition, SLAM, and rendering, and then sends the processed rendered video stream back to the XR glasses, where the final result is displayed. This solution ensures a high QoE while meeting the requirements for privacy protection.

4.2.3 Challenges

The main challenges of XR services are as follows: meeting the high-computing-capability requirements of XR tasks, satisfying extremely strict latency demands, and providing high-bandwidth support for the massive data transmission between the terminal and the network:

- **Ultra-low latency:** XR services are highly interactive and require extremely low latency. For example, VR experiences need latency below 20ms for basic functionality, below 15ms for comfort, and below 8ms for ideal performance [7]. Even with techniques like Asynchronous Timewarp (ATW), latency must be kept below 70ms [8]. This requires terminal-network coordination to reduce latency through QoS assurance, optimal path selection, and efficient computing frameworks.
- **High bandwidth:** XR devices have a wider field of view than traditional TVs, requiring higher resolutions, frame rates, and bit rates. For example, 8K VR video needs at least 260Mbps, while 12K/24K video requires over 1Gbps [7]. Terminal-network coordination must dynamically allocate sufficient bandwidth to support these high data rates
- **Ultra-high computing capability:** XR services involve processing large 3D models, complex encoding/decoding, and high-resolution, high-frame-rate rendering. For instance, 6-DoF volumetric video has significantly higher computational demands than traditional 2D video. This requires strong computing support from the network and efficient task management in terminal-network coordination.

Requirements for Terminal-Network Coordination:
To address these challenges, the network must provide high computing capability, low latency, and high bandwidth. It needs to support efficient communication between terminals and the network, offer QoS-assured network capabilities, and integrate computing services to deliver high-quality XR experiences.

4.3 Environmental Sensing through Terminal-Network Coordination

4.3.1 The Value of Applying Network Sensing Service in Intelligent Terminals

In future networks, a large number of intelligent terminals, such as autonomous vehicles, drones, delivery robots, and humanoid robots, rely on environmental sensing. However, these intelligent terminals face the following challenges in sensing and decision-making:

- **High terminal-side costs:** To enhance sensing accuracy and environmental adaptability, intelligent terminals are typically equipped with multiple sensing devices (such as cameras, LiDAR, and millimeter-wave radar) and require substantial AI computing capability. This leads to high overall device costs. For example, the AITO M9 is equipped with 192-line LiDAR, 11 high-definition cameras, and 3 millimeter-wave radars [9], and cost over 200,000 RMB.
- **Physical limitations of sensing:** The sensing of a single intelligent terminal is usually limited,

preventing it from obtaining a global view or dynamic environmental information. For example, the obstacle avoidance range of drones is typically only tens of meters, which cannot address issues like occlusions in 3D space or sensing of the environment at longer distances.

- **Lack of high-quality AI training data:** Highly intelligent terminals (such as embodied intelligent robots) heavily rely on high-quality, scenario-based training data. The acquisition of such data is a major bottleneck in the industry, hindering further improvements in intelligence levels.

The value of employing terminal-network coordination mechanisms is mainly reflected in the following aspects:

- **Reducing terminal-side costs:** Network sensing can provide real-time environmental maps, avoiding the expensive costs of surveying and updating. This allows delivery drones and autonomous delivery vehicles to operate without expensive sensing devices and AI computing capability, relying instead on sensing data provided by the network to achieve autonomous driving.
- **Expanding sensing range:** For intelligent terminals that require strong sensing capabilities (such as high-level autonomous vehicles and embodied intelligent robots), terminal-network coordination can provide global environmental assistance information to compensate for the limitations of individual sensing ranges.
- **Enhancing AI training efficiency:** For intelligent robots limited by a lack of scenario-based training data, terminal-network coordination can continuously collect multi-source sensing data. After fusion processing, this data forms a 3D spatial representation and is provided to intelligent robots for training and learning.
- **Multi-dimensional sensing capability:** Network sensing, through coordination between base stations and terminals, integrates various non-3GPP devices data sources such as video and radar, providing wide coverage, high-precision, and highly reliable sensing capabilities.

4.3.2 Use Case of Environmental Sensing with Terminal-Network Coordination

1) Scenario Description：

Taking the reduction of autonomous driving costs for unmanned delivery vehicles through network sensing services as an example (Figure 3), unmanned delivery vehicles are equipped with basic environment sensing devices, including cameras, active radars, and millimeter-wave radars, to support low-computing-capability local sensing capabilities such as traffic signal recognition and obstacle detection. The network sensing service provides real-time environmental maps as an alternative to the high-precision maps that were previously required to be pre-installed on the vehicle side. Logistics delivery companies provide global route planning (starting point, destination, travel route, etc.) and emergency escape mechanisms in abnormal situations (such as remote control for disentanglement) in the cloud.

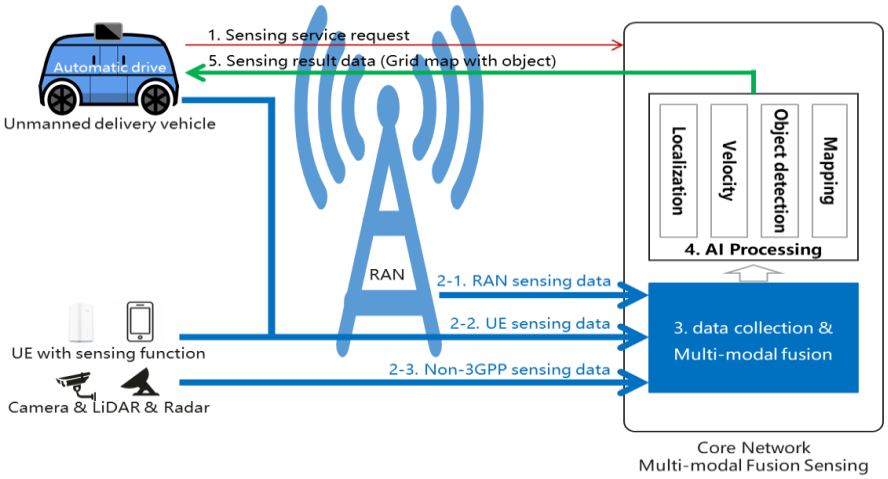

Figure 3 Use Case of Environmental Sensing with TNC

2) Potential Solutions for Applying Terminal-Network Coordination:

1. The unmanned delivery vehicle requests sensing services of electronic map for a specified area from the network.
2. Based on the request, the network decides which data to use (such as sensing base stations, sensing UEs, cameras, and radars) and notifies the corresponding sensing devices to start collecting data and transmitting it to the core network.
3. The core network receives and integrates these multi-source data, performing multidimensional processing through temporal and spatial alignment.
4. The core network processes the fused data (including localization, speed measurement, object detection, and mapping) to generate real-time environmental maps of dynamic and static objects.
5. The core network sends the processed results to the unmanned delivery vehicle. The vehicle combines the sensing data provided by the network with its own basic sensing capabilities to achieve autonomous navigation, obstacle avoidance, and other self-driving functions, thereby successfully completing the delivery task.

Based on the above process, the unmanned delivery vehicle can achieve high-level autonomous driving functions without relying on high-computing-capability devices, complex sensing equipment, expensive high-precision maps, or complex cloud-based AI model training.

4.3.3 Challenges

The main challenges faced network-assisted sensing for intelligent agents are as follows:

- **Precision of Network Sensing:** The precision of sensing can be described by multiple metrics, such as positioning accuracy, velocity measurement accuracy, and environmental map accuracy. According to the 6G use case "7.3 High-Resolution Terrain Map Use Case" approved by 3GPP SA1 [10], the 6G network is required to provide positioning accuracy of 0.1 meters, velocity measurement accuracy of 0.12 meters per second, and a resolution of 0.4 meters.
- **Real-time of Network Sensing:** When intelligent agent terminals request sensing services, the network must continuously provide real-time environmental maps. According to the 6G use case "7.3 High-Resolution Terrain Map Use Case" approved by 3GPP SA1, the maximum sensing service latency required by the 6G network is 50 ms, with a map refresh rate of less than 0.2 seconds.

- **Continuous and Seamless Coverage of Sensing Services:** When network sensing capabilities become an important source of environmental sensing data for intelligent agents, the network must provide continuous and seamless coverage of sensing services. This ensures that intelligent agents can autonomously and safely perform tasks in complex and dynamic public environments.
- **A direct way for interaction between terminal and network:** As many real time services need to be supported in 6G (e.g. autonomous driving), the timeliness of sensing information (e.g. the surrounding environment to a vehicle) is critical to terminal side. It is difficult to guarantee the timeliness of sensing information if the 6G network sensing information is always forwarded by application server to a terminal, instead, a direct interaction between terminal and 6G network will be much efficient and timeliness approved.

Requirements for Terminal-Network Coordination:

To address the above challenges, terminal-network coordination and network-assisted sensing need to:

- Integrate data characteristics and sensing capabilities from different sensing devices to achieve high sensing accuracy.
- Manage resources and scheduling for all sensing devices to enable the collection, transmission, storage, and sharing of multi-source sensing data.
- Utilize efficient AI algorithms to process the massive amounts of sensing data in real-time.

4.4 Intelligent Operation and Maintenance for Vertical Industries through Terminal-Network Coordination

4.4.1 The Value of Applying Network Sensing Computing Services in Intelligent Terminals

Under the future network architecture, the process of intelligent transformation in vertical industries is continuously accelerating. Taking the energy industry as an example, intelligent operation and maintenance is crucial for ensuring the efficient and stable operation of the energy system. A large number of intelligent terminals in the energy industry, such as smart meters, distributed energy collection devices, and inspection robots, have brought a lot of convenience to the operation and maintenance of the industry. However, the current intelligent operation and maintenance still faces many severe challenges:

- **High terminal costs:** In order to accurately collect energy data and adapt to complex and changeable environments, intelligent terminals need to be equipped with a variety of high-precision sensors such as high-precision current and voltage sensors and environmental monitoring devices, and also need to have strong local data processing capabilities. This undoubtedly makes the overall cost of the equipment rise significantly. Taking the intelligent wind turbine monitoring terminal of a large-scale wind farm as an example, with a variety of advanced sensors and high-performance processors it is equipped with, the cost can reach hundreds of thousands of yuan.
- **Physical limitations of sensing:** There are inherent limitations in the sensing range and capabilities of a single terminal, making it difficult to comprehensively obtain information and dynamic changes of the entire energy system. For example, environmental monitoring devices can only collect the short-distance environmental conditions at their locations, and cannot completely monitor the entire atmospheric environment where the wind turbines are located; inspection robots are prone to monitoring blind spots in the complex

environment of energy facilities, and it is difficult to detect the abnormal conditions of equipment in distant or blocked areas.

- **Lack of high-quality data:** The intelligent operation and maintenance in the energy industry highly depends on high-quality data based on actual scenarios, especially for the training of machine learning models used in predictive maintenance and fault diagnosis. However, currently, in the scenarios of new energy power generation, it is extremely difficult to effectively collect and label data on complex meteorological conditions and equipment operating conditions, which has become a key bottleneck restricting the improvement of the level of intelligent operation and maintenance.

In this context, the expected value of applying Terminal-Network Coordination in Intelligent Operation and Maintenance in Vertical Industries can be summarized as follows:

- **Reduce terminal costs:** Network sensing can provide real-time data on the operation status of the energy system. Intelligent terminals do not need to carry out complex and expensive data collection and processing work on their own. Taking the distributed energy collection terminal as an example, it does not need to be equipped with a high-cost data analysis module. Relying on the data analysis results provided by the network, it can effectively monitor energy production and transmission, thus reducing equipment costs.
- **Expand the sensing range:** For intelligent terminals with high requirements for sensing capabilities, such as the intelligent decision-making system of the energy dispatching center and the comprehensive monitoring system of large-scale energy facilities, terminal-network collaboration can integrate data from multiple terminals and the macro analysis provided by the network, providing comprehensive operation information of the energy system, making up for the insufficient sensing range of a single terminal, and helping to discover potential problems in advance.
- **Improve operation and maintenance efficiency:** Terminal-network collaboration can continuously collect multi-source sensing data such as smart meter data, equipment operation status data, and environmental monitoring data. After fusion processing, it can form a comprehensive status assessment of the energy system, providing strong support for intelligent operation and maintenance decisions. In this way, equipment failures can be predicted more accurately, maintenance plans can be arranged in advance, and the downtime of the energy system can be reduced.

4.4.2 Use Case of Intelligent Operation and Maintenance for Vertical Industries with Terminal-Network Coordination

1) Scenario Description:

Wind farms are usually located in the wild, and their operation and maintenance work faces many challenges. The operation and maintenance personnel of wind farms are equipped with basic detection tools, and the wind turbines in the wind farms are installed with basic condition monitoring devices, such as vibration sensors, temperature sensors, etc., which are used to initially sense the status of key components of the wind turbines, such as monitoring the temperature of the bearings and the vibration amplitude of the blades.

Despite the presence of basic monitoring equipment, the complex environment in the wild makes operation and maintenance quite difficult. For example, there are impacts of extreme weather, sensor failures, and the data is vulnerable to interference. The wind turbines are scattered, and it takes a long time to locate faults and reach the site. After the equipment ages, it becomes more difficult to detect potential hazards. The operation and maintenance standards for wind turbines of different models are different, and the operation and maintenance personnel need to keep learning and adapting, which poses certain pressure on both human and technical resources.

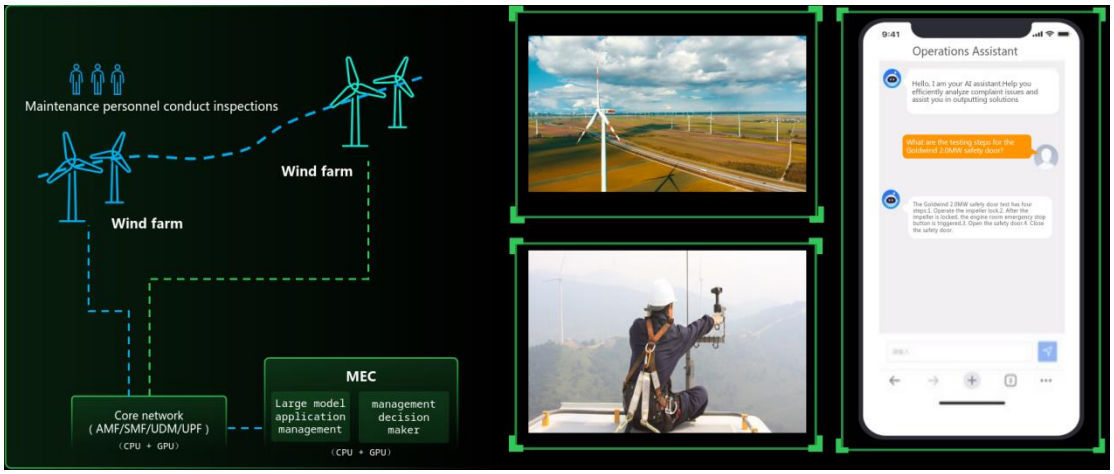

Figure 4 Use Case of Intelligent Operation and Maintenance for Vertical Industries with TNC

2) Potential Solutions for Applying Terminal-Network Coordination:

Provide low-latency communication services: Enable operation and maintenance personnel in the wild to interact in real time with various equipment in the wind farm and the remote management system. The base station is equipped with a communication protocol stack designed for future networks to ensure efficient data transmission and processing. Meanwhile, the core network, with the support of MEC (Multi-access Edge Computing), can manage large model applications as well as make scheduling and orchestration decisions for communication and computing resources.

Provide real-time operation and maintenance assistant services driven by large models: It can conduct in-depth analysis of the operation data of wind turbines and environmental data. For example, when a wind turbine experiences abnormal vibration, the operation and maintenance assistant can quickly analyze the vibration characteristics, combine historical data and meteorological conditions, accurately locate the cause of the fault, such as blade imbalance or gearbox wear, etc., and provide detailed maintenance suggestions and operation steps.

Provide sensing and monitoring services: During daily inspections, operation and maintenance personnel can use the wind farm operation and maintenance assistant application on mobile devices to query the status of wind turbines at any time, receive early warning information, and operate according to the guidance provided by the assistant, which greatly improves the intelligence and convenience of the operation and maintenance of the wind farm.

4.4.3 Challenges

Service resource negotiation and intelligent scheduling: Terminal-network coordination requires the realization of service resource negotiation and the intelligent management and scheduling of various network-side services through standardized service interfaces between terminals and the network. In the intelligent operation and maintenance of vertical industries, communication operators have to face different terminal devices and diverse business scenarios. How to ensure the effective establishment of these standardized interfaces, as well as the reasonable negotiation and intelligent scheduling of service resources in complex environments, are the main challenges.

Requirements for Terminal-Network Coordination

- Terminal-Network Coordination requires the establishment of efficient standardized service interfaces to achieve precise negotiation of service resources between terminals and the network, and conduct intelligent management of network-side services, so as to adapt to different terminal devices and diverse business scenarios, and ensure the reasonable allocation of service resources in complex environments.

5 Architecture and Key Technologies

In current 5G networks, the mechanisms for providing services from the network to terminals have many shortcomings. For example, with MEC, mobile operators typically collaborate with Over-The-Top (OTT) applications to pre-deployed application service software at the network edge. Terminal users then access these network-side services through application-layer interfaces for subscription, discovery, and invocation. However, this application-layer-based framework for terminal-cloud/edge coordination has the following issues:

- Pre-deployed application software at the network edge occupies fixed resources, leading to low resource utilization.
- Lack of coordination between application services and communication services in resource scheduling, which cannot guarantee the optimal service invocation path.
- The framework cannot decide whether to apply network services in terms of the dynamic network conditions.
- Terminal users must adapt to diverse application-layer interfaces to access network services, with no unified standard interface available. Additionally, each application layer independently develops capabilities like user authentication, service authorization, scheduling, security, and reliability, lacking a unified standardization mechanism.

In order to overcome such shortcomings and to address the challenges raised by the use cases in Clause 4, Terminal-Network Coordination architecture and key technologies are introduced, hoping to achieve low latency, high bandwidth and high reliability of services through model training, task decomposition and collaboration on the network side. **The key difference between the Terminal-Network Coordination mechanism in 6G and that in 5G lies in the fact that the network directly provides services to the terminal** (such as computing and sensing), empowering the terminal's native capabilities and breaking the boundaries of traditional network services. The goal of terminal-network coordination is to extend terminal capabilities through the network, provide comprehensive support to terminals, and lay a solid technical foundation for the diversity and continuous prosperity of application ecosystems. To achieve this goal, it is necessary to build a unified terminal-network coordination architecture, provide standardized application-layer interfaces, optimize communication and negotiation mechanisms between terminals and networks, and introduce intelligent management and scheduling technologies. Through these means, network-native services can be reasonably distributed on the terminal and network, enabling efficient empowerment of terminals by the network and significantly improving user service experience quality.

Therefore, this chapter proposes a terminal-network coordination architecture to support the implementation of the mechanisms. Through this architecture, service-oriented mechanisms, newly designed terminal-network communication protocols, intelligent scheduling frameworks for computational tasks, and collaborative computing service frameworks between terminals and networks, as well as related key technologies, can be further introduced. This lays the foundation for the comprehensive realization of terminal-network coordination.

5.1 Architecture

The key objective of the terminal-network coordination architecture is to establish a dynamic negotiation mechanism between terminals and the network. To achieve this, the following key capabilities are required:

- **Unified and Standardized Interface:** Provide a standardized network service invocation interface on the terminal side. Through this interface, terminal applications (users) can

initiate service requests using a unified API, reducing the complexity of adapting to different terminal types and operating system interfaces, thereby improving application development efficiency.

- **Unified Service Management and Scheduling:** Centralize the management of network-native services and resources. Integrate terminal, network-native, and external services. Manage and schedule them uniformly to optimize service invocation paths.
- **Service Coordinated Decision-Making:** Establish a service coordination mechanism between the terminal and network. Service tasks requested by terminal users can be dynamically allocated to either the terminal or network service nodes based on current conditions (such as terminal capabilities, network QoS, user plans, etc.), ensuring efficient resource utilization and optimal service performance.
- **Enhanced Security and Reliability:** The new terminal-network communication protocol carries service coordination messages. By leveraging the protocol's encryption and integrity protection capabilities, as well as the security of air-interface RRC signaling, the privacy of user information in terminal-network coordination messages is ensured. Additionally, low-latency and high-reliability performance of the terminal-network coordination mechanism is guaranteed.

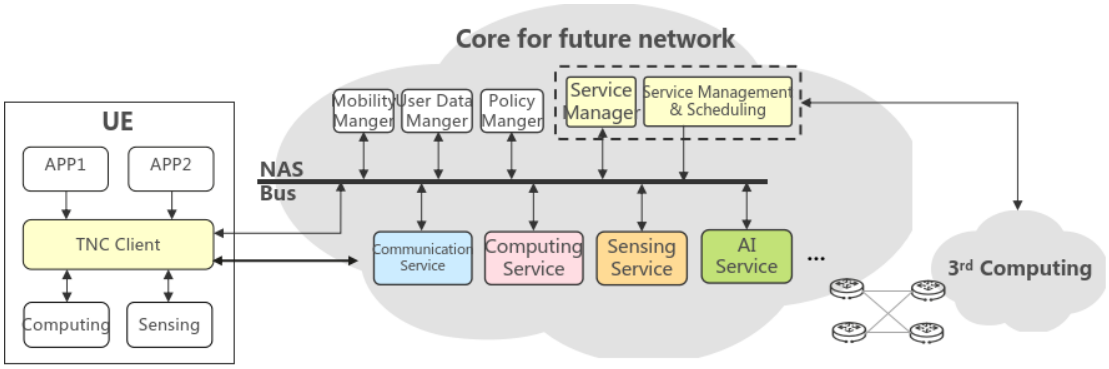

Figure 5 Overall Architecture Design of TNC

To realize the above capabilities, the overall architecture design of Terminal-Network Coordination (TNC) is shown in Figure 5. The TNC architecture includes several logical components, such as the TNC Client, Service Manager, Service Management & Scheduling, Computing Service, and Sensing Service. The functions of these components are as follows：

- **TNC Client:** Deployed on the terminal, the TNC Client provides standardized and service-oriented open APIs for upper-layer applications to trigger service subscription, request, and invocation by terminal users. It also includes a set of NAS services responsible for NAS signaling interaction with the network to complete service subscription, registration, discovery, and scheduling processes. Through coordination with the network, the TNC Client enables dynamic switching between local and network services for terminal applications, ensuring continuous user experience.
- **Service Manager:** The Service Manager is responsible for service management. It maintains registration information for terminal-side services, network-native services, and external services, including topology, resource availability, and service performance assessment, while monitoring service status in real-time. The Service Manager provides interfaces for Service Management & Scheduling to obtain service information and status change events, enabling service management and scheduling. It also interacts with the TNC Client to complete service subscription and distribution processes for terminal users.
- **Service Management & Scheduling:** The key function of Service Management & Scheduling is service logic management and resource scheduling. It selects appropriate service nodes based on user demands and service information status, and completes resource allocation

and service instantiation. By invoking the Service Manager's interfaces, Service Management & Scheduling generates the required service information and status-change events to efficiently complete the management and scheduling process

- **Network-Native Service Units — Computing Service, Sensing Service, etc.:** These service units provide underlying infrastructure resources (such as heterogeneous computing capability, storage, and network resources) and upper-layer application services (such as AI inference, rendering, and sensing). They register topology information, available resources, and service performance assessment information with the Service Manager, and respond to service allocation requests from Service Management & Scheduling to complete resource allocation and service instantiation. The Computing and Sensing Services are also responsible for creating computing and sensing sessions, and providing runtime status, duration, and underlying resource consumption statistics for service business analysis and billing.

Through the collaboration of these logical components, the complete process from service registration, management to execution is realized, effectively enhancing the system's service-oriented and intelligent levels.

5.2 Service-Oriented Mechanism of Terminal-Network Coordination

The purpose of the terminal-network coordination service-oriented mechanism is to enable the network to flexibly empower terminals with various services (including connectivity, computing, and sensing services) on demand. This mechanism supports service registration and discovery between the terminal and network sides, schedules optimal service nodes for terminal users, establishes high-quality QoS communication channels, and meets users' personalized service needs. The specific process of the terminal-network coordination service-oriented mechanism is shown in Figure 6.

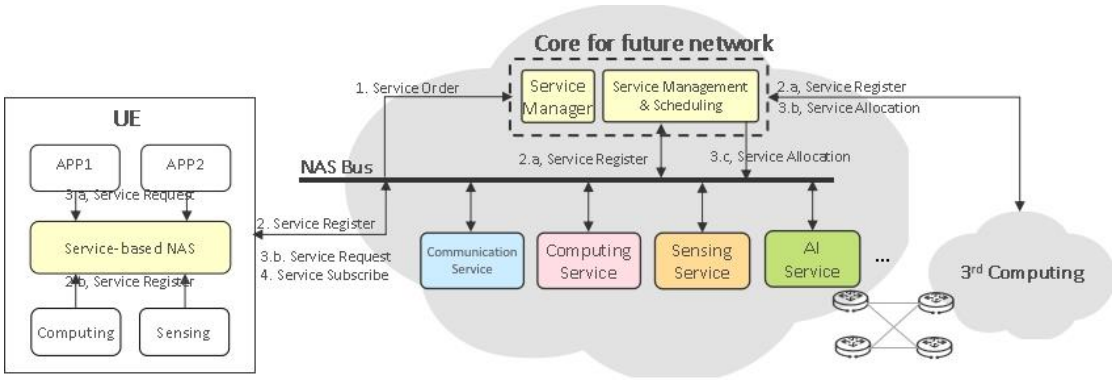

Figure 6 Specific Process of TNC Service-Oriented Mechanism

1. Dynamic Service Subscription：

Terminal applications initiate subscription requests to the network through service-oriented APIs to obtain specific network services (e.g., computing or sensing services). These services can be operated by the carrier or user-defined. The subscription request includes user identity information, location, subscription preferences, service customization configurations, and may also carry user privacy and security requirements (e.g., specifying whether the service can be shared with other users). After completing user authentication, the network assigns a service ID,

generates subscription data and policies, and sends a subscription response back to the terminal user, including subscription results, security credentials, and service ID.

2. Diverse Service Registration

Services that need to be registered with the network include network-native services, third-party external services, and terminal services:

- **2.a Carrier Network Services and External Third-Party Services:** These services register with the network, providing information such as service node topology, available service resources, service software, service performance assessment, and service connection QoS.
- **2.b Terminal-Side Services:** Terminal local services invoke the native NAS service open API provided by the terminal-network coordination architecture to register with the network, including available terminal resources and service information.

The network manages this service information and, when users request services, allocates service tasks flexibly based on resource conditions on both the terminal and network, user's requirements and network transmission QoS to fully utilize terminal-network resources and ensure the best user experience.

3. Unified Service Request

- **3.a Terminal Application's Service Request:** Terminal applications initiate service requests to the network through service APIs, providing service identifiers and required resource information.
- **3.b Service Request Transmission:** The terminal sends the service request to the network via the NAS protocol.
- **3.c Network Service Request Processing:** The network processes the user's service request, analyzing comprehensively based on user location, service requirements (including service type, resource requirements, KPI requirements), available resources on both the terminal and network, network transmission QoS, and privacy and security requirements. The network then decides on the optimal management and deployment of services between the terminal and network and feeds back the service management results to the terminal.

4. Terminal Subscription to Network Services

The network can also offer service subscriptions to terminals. Terminal applications initiate subscription requests to the network side through service APIs, attaching service filtering conditions (such as desired service types, service provision areas) and possibly including user privacy requirements or preferences. The network side screens available services based on these conditions and returns a list to the terminal.

A typical scenario of service subscription is as follows: Users can subscribe to a service in advance, even if it is not yet enabled in their current area. When the user enters the service area, the network side will proactively notify and push the service to the user for use.

5.3 Service-Oriented NAS Communication Protocol Design

To achieve secure and efficient negotiation between terminals and the network, the terminal-network coordination architecture introduces a service-oriented NAS protocol mechanism. This mechanism carries key messages, including service subscription, service registration, service discovery, service session management, and service status synchronization, as well as user privacy information (such as user identifiers, location information, personalized preferences, and security credentials) to ensure security, low latency, and high reliability during transmission.

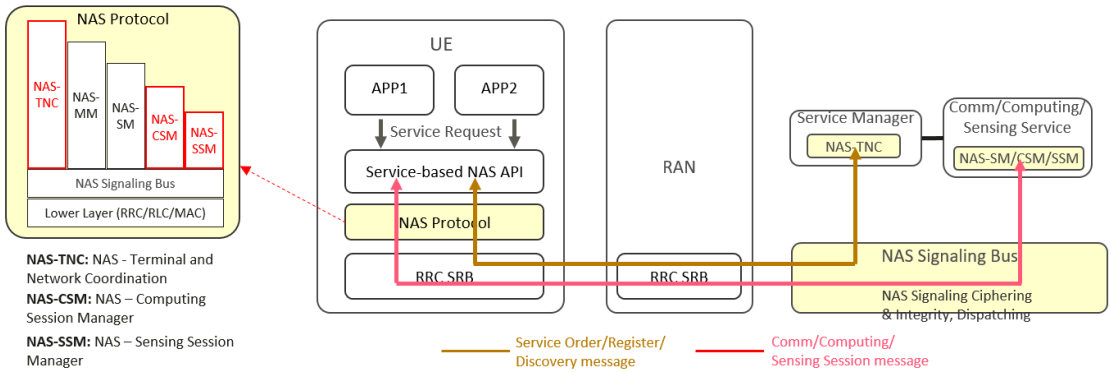

Figure 7 New NAS Protocol Proposed by TNC

The new NAS protocol proposed by terminal-network coordination is shown in Figure 7. This protocol enhances the current 5G NAS architecture in the following ways:

1) **Decoupling of NAS-MM and Unified Signalling Distribution:** NAS-MM will no longer be responsible for distributing other NAS services. Instead, a NAS Signalling Bus layer is introduced to centrally manage signalling distribution between terminals and the network. This adjustment decouples NAS-MM from signalling distribution, allowing it to focus on its core function of mobility management.
2) **New NAS Services in the Protocol Layer:** In addition to the existing 5G NAS-MM and NAS-SM services, the NAS protocol layer on both the terminal and network sides introduces three new services: NAS-TNC, NAS-CSM, and NAS-SSM.
   a) **NAS-TNC** is responsible for the interaction process of service-oriented NAS signalling between terminals and the network.
   b) **NAS-CSM** handles NAS signalling for the establishment, update, and deletion of computing sessions, ensuring QoS for terminal users when invoking network computing services and establishing user-plane communication channels.
   c) **NAS-SSM** manages NAS signalling for the establishment, update, and deletion of sensing sessions, providing QoS-guaranteed user-plane communication channels for terminal users when invoking network sensing services.
3) **Open Service-Based NAS API:** The introduction of the Service-Based NAS API allows terminal applications (users) to conveniently initiate processes such as service subscription, service discovery, service subscription, and service status perception. This standardized interface not only simplifies the process of initiating service requests by upper-layer applications (users) but also significantly reduces the complexity of adapting to various terminal types and operating system interfaces.

5.4 Intelligent Service Management and Scheduling Mechanism

When terminal applications (users) initiate task requests, they often need to invoke multiple services simultaneously. To achieve efficient resource utilization and ensure the best user experience, there is an urgent need for an intelligent management and scheduling mechanism that can optimally allocate tasks between service nodes on the terminal and network. The key functions of the service management and scheduling mechanism include **service logic management, service resource management, and service path scheduling.** As shown in Figure 8, the logical process is as follows: After receiving the Service Request from the terminal user, the Service Management & Scheduling on the network side first performs service logic management,

i.e., it determines the required service types and their execution logic based on the user request.

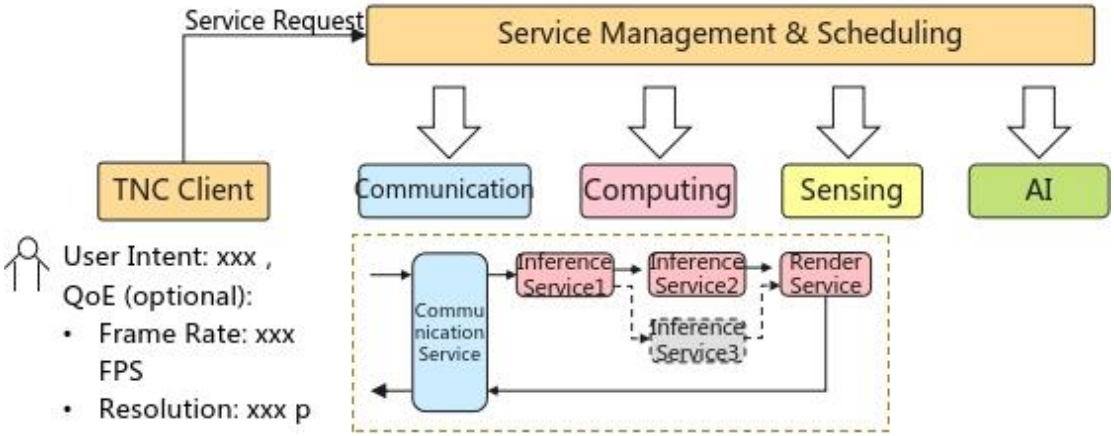

Figure 8 Service Management and Scheduling Mechanism of TNC

Service Logic Management: After receiving the Service Request from the terminal, the Service Management & Scheduling on the network side can use AI-based large models to identify key info in user intents, thereby understanding the intent and managing the service execution logic (e.g., from Communication Service to Inference Service 1, then to Inference Service 2, and finally to Rendering Service). This approach differs from traditional methods that rely on predefined rule tables or intent tables. Instead, it dynamically recognizes and maps service logic through AI understanding, overcoming the limitations of traditional methods in handling ambiguous intents and achieving higher accuracy.

Service Resource Management: Traditional management methods typically rely on predefined resource templates and divide user experience requirements into several SLA levels. The manager extracts parameters from the user intent request and selects a matching resource template. Specifically, resource requirements include service resource types (e.g., connectivity, computing, or sensing), processor types (e.g., CPU, GPU, NPU), and detailed resource parameters (e.g., computing power requirement of xx FLOPS, memory requirement of xx MB, communication bandwidth requirement of xx Mbps).

Traditional template selection, however, may lead to insufficient or wasted resources. The terminal-network coordination architecture uses AI-based models to train a resource knowledge base for various services and adjusts it in real-time based on resource consumption data during service runtime. This method continuously improves the resource knowledge base. Additionally, AI inference can deeply understand personal preferences (e.g., QoE parameters) embedded in user intents, thereby more accurately assessing resource requirements for various services, achieving more efficient resource allocation, and providing a better user experience.

Service Path Scheduling: The terminal-network coordination architecture introduces an integrated scheduling mechanism for communication connections and business services. By considering multiple factors such as user location, service/communication resource requirements, privacy and security needs, and latency experience, it uniformly selects the most suitable service nodes, interconnection nodes, and terminal-network communication nodes to manage the optimal service execution path.

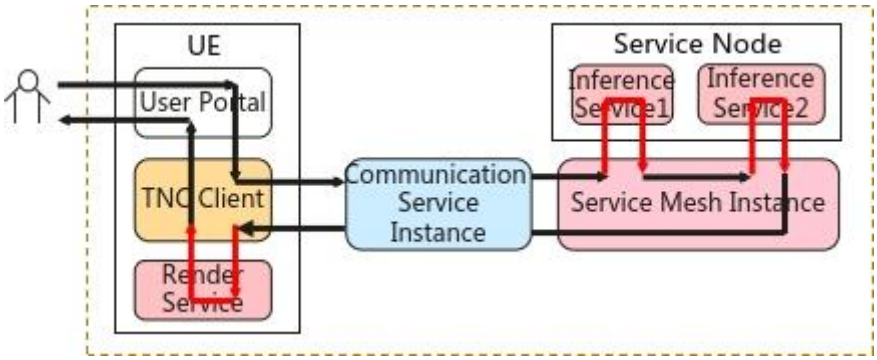

Figure 9 Service Path Scheduling for TNC

For example, as shown in Figure 9, inference services may require significant computing power. If the terminal-side computing resources are insufficient or the latency is high, it could negatively impact user experience. Therefore, the manager may allocate these inference services to network-side service nodes for processing. In contrast, rendering services, which have lower computing power requirements, can be completed on the terminal side with low latency. Additionally, since the transmission of rendering results may consume significant bandwidth, the manager decides to execute this service locally on the terminal side. For services allocated to network-side nodes, the manager schedules the optimal service path based on privacy and security, resource requirements, and the principle of minimizing end-to-end latency. In the future, the management and scheduling scheme can further incorporate factors such as cost, energy efficiency, and user personal preferences to enhance user experience and operational efficiency.

5.5 Computing Service Framework

The terminal-network coordination computing service framework is designed to construct, manage, and optimize computing resources and services for terminal-network coordination. It provides diverse computing capabilities, including general-purpose computing, intelligent computing, and super computing, and connects them through a computing capability integrated network. Supported by various computing related technologies, the framework unifies the management of heterogeneous computing capability outputs. Moreover, it integrates deeply with technologies such as AI and block-chain, encapsulating computing, storage, and networking resources into a unified package, which is then provided to users in the form of services (e.g., via APIs).

However, directly migrating the runtime environment of terminal applications to edge or cloud computing platforms may face challenges such as incompatibility of heterogeneous computing capability, difficulty in managing remote computing resources, and network congestion caused by imbalanced supply and demand of computing capability. Therefore, the terminal-network coordination computing service framework should possess the following key capabilities:：

1) **Cross-Heterogeneous Platform Migration:**
   The computing service framework should abstract different heterogeneous computing resources through techniques such as resource pooling, standard run times, and unified programming models. This allows applications to migrate smoothly between different computing platforms, achieving the goal of "one set of code, universal network compatibility".
2) **"Function as a Service (FaaS)":**
   Essentially, users expect to obtain a runtime environment for their applications, focusing primarily on business logic without needing to, or with minimal attention to, technical

details (e.g., performance, scalability, high availability). The computing service framework should provide service interfaces that allow users to deploy their programs directly onto the platform. Resources are dynamically allocated upon event triggers to handle tasks and released after processing is complete, thus realizing "Function as a Service" (FaaS).

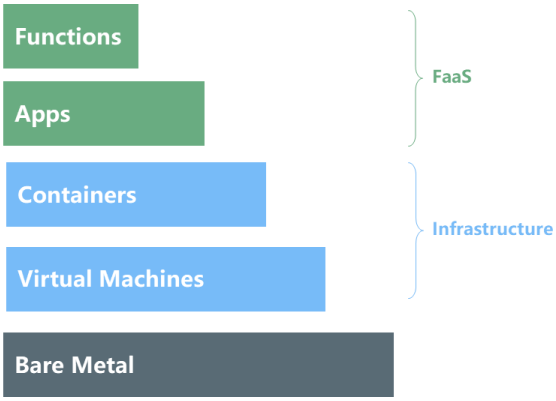

Figure 10 FaaS

3) Native Distributed Architecture:

The computing service framework should offer micro services-based, dynamic, and distributed computing capability. It should decompose traditional, stateful, coarse-grained applications into micro services and fine-grained components, which are then deployed across the most suitable computing resources. By fully leveraging a distributed architecture, the framework can maximize the service capabilities of network-side computing resources, enhance computational performance, and improve resource utilization efficiency.

These key capabilities will collectively drive the efficient operation of the terminal-network coordination computing service framework, delivering a more flexible, efficient, and reliable computing service experience for users.

6 The Value of Network Computing Enabling Terminal-Network Coordination

Today, mobile networks, while providing massive basic mobile connections, are also beginning to offer capabilities beyond connectivity, such as network-native computing and sensing, as their technological capabilities improve. Through a service-oriented terminal-network coordination mechanism, efficient terminal-network interoperability protocols, intelligent service management and scheduling, and more advanced computing service frameworks, the network and terminals achieve deep integration. This empowers terminals with prosperous applications and enables new capabilities for the 6G network.

1) Low-Latency and Efficient Services through Integrated Management and Scheduling

When providing terminal computing or sensing services, mobile networks leverage their proximity to terminals to offer low-latency, high-bandwidth services for applications requiring rapid responses, such as AR, VR, and autonomous driving. For example, in providing low-latency network-native computing services, the network considers the current network quality (bandwidth and latency) and the status of deployed computing services/resources. Based on user requests, it offers an optimal integrated management plan that meets both computing needs and network conditions. Through integrated scheduling, allocation, and routing, the network achieves synergy between connectivity and computing/sensing services, ensuring end-to-end optimization of the service path. This integrated scheduling mechanism avoids the issue of suboptimal end-to-end path utilization when applications independently build terminal-cloud/edge coordination mechanisms, thereby providing users with higher-quality real-time services.

2) Unified Service Access Platform Across Vendors and Ecosystems

Mobile networks build a unified service access platform across vendors and ecosystems through standardized interfaces and protocols. Under the terminal-network coordination mechanism, this platform provides standardized interface services for diverse applications to achieve unified access to network-inherent computing and sensing services.

On the other hand, by extending the standardized NAS signalling interactions between terminals and networks through a service-oriented mechanism, mobile networks can avoid compatibility issues between different vendors' independent ecosystems. This approach offers significant advantages in security, low latency, and reliability:

a) The interaction of messages carrying user privacy information between terminals and the network side is carried over NAS signalling, leveraging the encryption and integrity protection capabilities inherent in the NAS protocol stack of mobile communication networks. This better protects user privacy and prevents malicious tampering.

b) NAS messages are transmitted over a signalling channel rather than a mobile data connection channel, ensuring lower latency and higher reliability for signalling.

In summary, facing a diverse range of applications and terminal roles, the operators, as neutral players, can provide unified and public terminal-network coordination services. This breaks through the limitations of single-vendor ecosystems and promotes multi-party collaboration and innovation.

3) Enabling Diverse Future Terminal Forms and Supporting Rapid Application Innovation

Mobile networks provide neutral and real-time edge capabilities for small and medium-sized companies that lack robust cloud computing support. For these enterprises, building their own cloud platforms is not only costly but also challenging for scalable operations. Compared to cloud services offered by OTT providers, mobile networks' edge cloud services cover a broader area and offer lower latency, effectively addressing the shortcomings of OTT

vendors in edge cloud coverage.

For example, many small and medium-sized AR glasses companies lack sufficient cloud infrastructure. By providing distributed edge computing resources, mobile networks offer strong technical support, enabling these companies to quickly develop new products and efficiently bring them to market without building complex computing centers.

Through its native terminal-network coordination mechanism, mobile networks provide neutral, secure, real-time, reliable, and sustainable diverse service capabilities. This model not only helps enterprises reduce technical barriers and operational costs but also provides strong driving force for the continued prosperity of future AI and related industrial ecosystems.

7 Conclusion and Future Work

To build the capability for intelligent terminal-network coordination at the beginning of 6G, it is necessary for the telecommunications industry, including operators, terminal manufacturers, OTT application providers, and communication equipment vendors, to work together to promote industrial reform and build the core competitive edge of the future telecom network technology market:

- **Strengthen standard guidance and build ecosystem:** Leveraging the decades of standardization experience and cross-ecosystem core competitiveness of standards organizations such as 3GPP and ETSI in the mobile communications field, we should actively guide and promote the formulation and implementation of 6G standards. Through standardization efforts, we can reinforce industry confidence, use the important opportunity of network evolution and generational turnover, and exert the influence of standards to lead the entire ecosystem toward the target direction, continuously enhancing the competitiveness of the network.
- **Seize the opportunity of AI innovation to achieve synchronized development:** AI innovation poses higher requirements for the network, such as high bandwidth, low latency, and high reliability. The 6G network should fully utilize AI technology to empower autonomous networks and improve the intelligence level of the network. At the same time, the 6G network should also provide strong network service support for AI innovation applications and terminal capabilities, explore and develop in synchronization with AI technology, and form a virtuous cycle of mutual promotion and common progress.

As the next generation in the communication industry, 6G carries the responsibility of technological innovation, product replacement, and consumption upgrade. It is a key opportunity for the information and communication industry ecosystem to achieve a transformative leap. Guided by demand scenarios, addressing the pain points of 5G and serving emerging businesses are the core directives for 6G development.

Integrating emerging elements such as AI, computing power, and sensing is the necessary path for 6G's innovative development. These elements will merge to build a powerful 6G network, providing users with unprecedented service experiences. The Network Computing Enabling Terminal-Network Coordination of 6G requires the joint efforts of the entire industry to seize the opportunities of the new generation of mobile communications and achieve a win-win situation.

8 References

[1] Qualcomm, "The future of AI is hybrid," 2023.

[2] Apple Security Research, "Private Cloud Compute Security Guide," 2024. [Online]. Available: https://security.apple.com/documentation/private-cloud-compute.

[3] Meta, "Introducing quantized Llama models with increased speed and a reduced memory footprint," [Online]. Available: https://ai.meta.com/blog/meta-llama-quantized-lightweight-models/.

[4] M. Agarwal, A. Qureshi and N. Sardana, "LLM Inference Performance Engineering: Best Practices," databricks, [Online]. Available: https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices.

[5] Qualcomm, "Snapdragon AR2 Gen 1 Platform," [Online]. Available: https://www.qualcomm.com/products/mobile/snapdragon/xr-vr-ar/snapdragon-ar2-gen-1-platform.

[6] Meta, "Introducing Orion, Our First True Augmented Reality Glasses," Meta, [Online]. Available: https://about.fb.com/news/2024/09/introducing-orion-our-first-true-augmented-reality-glasses/.

[7] H. iLab, "Cloud VR Network Solution White Paper," [Online]. Available: https://www-file.huawei.com/-/media/corporate/pdf/ilab/2018/cloud_vr_network_solution_white_paper_2018_en_v1.pdf.

[8] H. Z. E. Q. M. China Mobile, "GTI XR NetworkTechnologyWhite Paper," 2023.

[9] A. M9, "AITO M9," [Online]. Available: https://aito.auto/model/m9/.

[10] 3GPP, "S1-2444696 TR22.870_Clean post Orlando.docx".